Universidad de Alcalá
COMISIÓN DE ESTUDIOS OFICIALES
DE POSGRADO Y DOCTORADO

ACTA DE EVALUACIÓN DE LA TESIS DOCTORAL

(FOR EVALUATION OF THE ACT DOCTORAL THESIS)

Año académico (*academic year):* 2017/18

DOCTORANDO *(candidate PHD)*: **NOGALES MOYANO, ALBERTO**
D.N.I./PASAPORTE (*Id.Passport*): ******0222F**
PROGRAMA DE DOCTORADO *(Academic Committee of the Programme)*: **D442-INGENIERÍA DE LA INFORMACIÓN Y DEL CONOCIMIENTO**
DPTO. COORDINADOR DEL PROGRAMA *(Department)*: **CIENCIAS DE LA COMPUTACIÓN**
TITULACIÓN DE DOCTOR EN *(Phd title)*: **DOCTOR/A POR LA UNIVERSIDAD DE ALCALÁ**

En el día de hoy 23/07/18, reunido el tribunal de evaluación, constituido por los miembros que suscriben el presente Acta, el aspirante defendió su Tesis Doctoral **con Mención Internacional** (*In today assessment met the court, consisting of the members who signed this Act, the candidate defended his doctoral thesis with mention as International Doctorate*), elaborada bajo la dirección de (*prepared under the direction of*) ELENA GARCÍA BARRIOCANAL // MIGUEL ÁNGEL SICILIA URBAN.

Sobre el siguiente tema *(Title of the doctoral thesis)*: *A STRUCTURAL AND QUANTITATIVE ANALYSIS OF THE WEBOF LINKED DATA AND ITS COMPONENTS TO PERFORM RETRIEVAL DATA*

Finalizada la defensa y discusión de la tesis, el tribunal acordó otorgar la CALIFICACIÓN GLOBAL[1] de **(no apto, aprobado, notable** y **sobresaliente)** *(After the defense and defense of the thesis, the court agreed to grant the GLOBAL RATING (fail, pass, good and excellent)*: SOBRESALIENTE

Alcalá de Henares, a 23 de Julio de 2018

Fdo. *(Signed)*: Fabio Sartori Fdo. *(Signed)*: Fdo. *(Signed)*: Alvaro J. Garcia

FIRMA DEL ALUMNO (candidate's signature),

Fdo. *(Signed)*: Alberto Nogales

Con fecha 26 de septiembre de 2018 la Comisión Delegada de la Comisión de Estudios Oficiales de Posgrado, a la vista de los votos emitidos de manera anónima por el tribunal que ha juzgado la tesis, resuelve:

☒ Conceder la Mención de "Cum Laude"
☐ No conceder la Mención de "Cum Laude"

La Secretaria de la Comisión Delegada

[1] La calificación podrá ser **"no apto" "aprobado" "notable"** y **"sobresaliente"**. El tribunal podrá otorgar la mención de ~~cum laude~~ si la calificación global es de sobresaliente y se emite en tal sentido el voto secreto positivo por unanimidad. (*The grade may be "**fail**" "**pass**" "**good**" or "**excellent**". The panel may confer the distinction of "cum laude" if the overall grade is "Excellent" and has been awarded unanimously as such after secret voting.).*

INCIDENCIAS / OBSERVACIONES:
(Incidents / Comments)

Universidad de Alcalá

COMISIÓN DE ESTUDIOS OFICIALES
DE POSGRADO Y DOCTORADO

En aplicación del art. 14.7 del RD. 99/2011 y el art. 14 del Reglamento de Elaboración, Autorización y Defensa de la Tesis Doctoral, la Comisión Delegada de la Comisión de Estudios Oficiales de Posgrado y Doctorado, en sesión pública de fecha 26 de septiembre, procedió al escrutinio de los votos emitidos por los miembros del tribunal de la tesis defendida por *NOGALES MOYANO, ALBERTO*, el día 23 de julio de 2018, titulada *A STRUCTURAL AND QUANTITATIVE ANALYSIS OF THE WEBOF LINKED DATA AND ITS COMPONENTS TO PERFORM RETRIEVAL DATA*, para determinar, si a la misma, se le concede la mención "cum laude", arrojando como resultado el voto favorable de todos los miembros del tribunal.

Por lo tanto, la Comisión de Estudios Oficiales de Posgrado **resuelve otorgar** a dicha tesis la

MENCIÓN "CUM LAUDE"

Alcalá de Henares, 4 de octubre de 2018
EL VICERRECTOR DE INVESTIGACIÓN Y TRANSFERENCIA

UNIVERSIDAD DE ALCALÁ · VICERRECTORADO DE INVESTIGACIÓN Y TRANSFERENCIA

F. Javier de la Mata de la Mata

Copia por e-mail a:
Doctorando: NOGALES MOYANO, ALBERTO
Secretario del Tribunal: SALVADOR SÁNCHEZ ALONSO.
Directores de Tesis: ELENA GARCÍA BARRIOCANAL//MIGUEL ÁNGEL SICILIA URBAN

Universidad de Alcalá

ESCUELA DE DOCTORADO
Servicio de Estudios Oficiales de Posgrado

DILIGENCIA DE DEPÓSITO DE TESIS.

Comprobado que el expediente académico de D./Dª Alberto Nogales Moyano reúne los requisitos exigidos para la presentación de la Tesis, de acuerdo a la normativa vigente, y habiendo presentado la misma en formato: ☒ soporte electrónico ☐ impreso en papel, para el depósito de la misma, en el Servicio de Estudios Oficiales de Posgrado, con el nº de páginas: 141 se procede, con fecha de hoy a registrar el depósito de la tesis.

Alcalá de Henares a 15 de junio de 2018

Purificación Reviejo

Fdo. El Funcionario

Universidad de Alcalá

JOSÉ JAVIER MARTÍNEZ HERRÁIZ, Coordinador de la Comisión Académica del Programa de Doctorado en ING. INFORMACIÓN Y CONOCIMIENTO

INFORMA que la Tesis Doctoral titulada A STRUCTURAL AND QUANTITATIVE ANALYSIS OF THE WEBOF LINKED DATA AND ITS COMPONENTS TO PERFORM DATA RETRIVAL, presentada por D/Dª ALBERTO NOGALES HOYANO, bajo la dirección del / de la Dr/a. MIGUEL A. SICILIA Y ELENA GARCIA, reúne los requisitos cientfficos de originalidad y rigor metodológicos para ser defendida ante un tribunal. Esta Comisión ha tenido también en cuenta la evaluación positiva anual del doctorando, habiendo obtenido las correspondientes competencias establecidas en el Programa.

Para que así conste y surta los efectos oportunos, se firma el presente informe en Alcalá de Henares a 20 de feb. de 2018

UNIVERSIDAD DE ALCALÁ · FACULTAD DE CIENCIAS DE LA COMUNICACIÓN

Fdo.:

Universidad de Alcalá

D. Miguel A. Sicilia y Dña. Elena García Barriocanal, directores de la tesis doctoral titulada "A structural and quantitative analysis of the Web of Linked Data and its components to perform data retrieval" y llevada a cabo por D. Alberto Nogales Moyano,

HACEN CONSTAR que la citada Tesis Doctoral ha sido realizada por compendio de artículos, reuniendo los requisitos exigidos a este tipo de tesis, así como los requisitos científicos de originalidad y rigor metodológicos para ser defendida ante un tribunal. Esta Comisión ha tenido también en cuenta la evaluación positiva anual del doctorando, habiendo obtenido las correspondientes competencias establecidas en el Programa.

Para que así conste a los efectos del depósito de la tesis, se firma en Alcalá de Henares a 13 de junio de 2018

Fdo.: Miguel A. Sicilia

Fdo.: Elena García Barriocanal

Universidad de Alcalá

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN,

ESCUELA POLITÉCNICA

Doctoral PhD

A structural and quantitative analysis of the Web of Linked Data and its components to perform data retrieval.

PhD candidate: Alberto Nogales

Supervisors: Elena García Barriocanal / Miguel Ángel Sicilia Urbán

A mis padres y a mi hermana por la ayuda durante estos años.

A mis amigos por el apoyo.

- This epic problem is not a problem for me –

Acknowledgments

I would like to thank all the people involved in the process of this thesis. First of all, my parents and my sister for all the help they have given me during this process and since I started to work as a researcher. The rest of my family for being interested in the advances of this work.

Also, I would like to thank everybody at IERU research group. In special, Miguel Ángel Sicilia Urbán who has guided me during all the process and has helped during the hardest moments. It has also been very useful the help of Elena García Barriocanal and Salvador Sanchez Alonso and the rest of my workmates at the lab. Special mention to David Martín and Juan Ruiz, for sharing a lot of time together working side by side.

I don't want to forget other people who has been a support during the last 5 years. My "brothers" from school, who have spent time with me since we were two years old. My friends from Medicine with whom I have been hours and hours studying. NZT friend, who have been hearing my boring speeches about what a PhD consists of. My friends in Madrid: roommates, Asturies posse, russkies, workmates at Freeway, Esteban, Amalio, Rodri, Catalina, David, Victor and Hector. Last but not least, the guys from "el último mono" for cheering me up the day it seems like everything was over.

Abstract

This research consists of a quantitative and structural analysis of the Web of Linked Data to improve the prospects for data retrieval. The Web of Linked Data arose when companies and organizations started to publish data sources that could be openly accessed by Web users. These datasets had different mechanism of access and formats, so Tim Berners Lee proposed the four principles for publishing and interlinking structured data on the Web.

In order to obtain quantitative metrics of the Web of Linked Data, statistical techniques are applied. In the case of the structural analysis Social Network Analysis (SNA) is used. SNA is the process to study of the relations of link structures applying graph and network theory. Nodes and edges form these kinds of structures. The nodes represent the actors and the edges represent the relations between them.

To have a snapshot of the Web of Linked Data in order to make the analysis, we started from the Linked Open Data (LOD) cloud diagram. This is an online catalogue of datasets whose information have been published using Linked Data techniques. These sets of data have been created by companies, organizations and individuals of the Open Data Movement interested in opening their own information so regular users could work with them. The datasets are published in a language called Resource Description Framework (RDF), which creates links between them, so information could be reused.

The aim of obtaining a quantitative and structural analysis of the Web of Linked Data is to improve data retrieval. Having an in-depth idea of the structure and the characteristics of LOD, it is possible to enhance the use of its data. In future works, users' searches could be faster and more accurate. For that purpose, we will take advantage of the use of the vocabulary Schema.org and the project LOV (Linked Open Vocabularies).

Schema.org is a set of tags whose purpose is that Webmasters could mark-up their own Websites with microdata. Microdata is used to help search engines and other Website tools to better understand the information contained in the Websites. LOV is a catalogue to register all the vocabularies used by the datasets

from the Web of Linked Data. Its aim is to provide an easy access to the vocabularies.

In this research, we are reporting a study on the mechanisms that may enhance data retrieval from the Web of Linked Data using the previous resources and ontology matching techniques. These techniques aim to map terms from two different sources and obtain which of them are common to both sources. In our case, first we are mapping Schema.org with LOV, and then LOV with the Web of Linked Data.

A network analysis of LOV has also been reported. The aim of this analysis is to obtain a quantitative and structural insight of LOV. Knowing this we can conclude which are the most popular vocabularies or if they are specialized in a particular field. This can be used to filter datasets or reuse information.

The findings show different issues. In the case of the structure of the Web of Linked Data, it is concluded that is compact and the distance between nodes is low. Also, it has been checked that it follows the bow-tie theory and the most important datasets are WordNet 2.0 and DBpedia. Taking into account the analysis made in LOV, the following conclusions have been extracted. The vocabularies are not specialized in a particular field and there is no dominant scope. Also, the most popular vocabularies correspond to standards of the Semantic Web or that used to model other vocabularies like RDF, OWL (Web Ontology Language) or SKOS (Simple Knowledge Organization System). Finally, with the mappings between Schema.org, LOV and the Web of Linked Data, we have developed two use cases in data retrieval. The first let the users enrich Websites with information obtained from the datasets of LOD. The other use case consists of extending ontologies with new classes and properties of Schema.org. Another independent use case presented in this research as additional contribution consist of retrieving information from Google Scholar and aggregating it to sources that storage scientific knowledge like VIVO and CERIF.

Resumen

Esta investigación consiste en un análisis cuantitativo y estructural de la Web of Linked Data con el fin de mejorar las perspectivas para la búsqueda de datos. La Web of Linked Data surgió cuando compañías y organizaciones empezaron a publicar repositorios de datos abiertos a los que los usuarios podían acceder. Estos conjuntos de datos tenían diferentes mecanismos de acceso y formatos, por lo que Tim Berners Lee propuso los cuatro principios para publicar e interconectar datos estructurados en la Web.

Con el objetivo de obtener métricas cuantitativas de la Web of Linked Data, se aplicarán técnicas estadísticas. En el caso del análisis estructural se usará un Análisis de Redes Sociales (ARS). ARS es un proceso para estudiar las relaciones de estructuras sociales aplicando teorías de grafos y redes. Estas estructuras se forman por nodos y arcos. Los nodos representan los actores y los arcos las relaciones entre ellos.

Para tener una idea de la Web of Linked Data poder hacer un análisis de su estructura, empezaremos con el diagrama de la Linking Open Data (LOD) cloud. Éste es un catálogo online de datasets cuya información ha sido publicada usando técnicas de Linked Data. Estos sets de datos han sido creados por compañías, organizaciones y personas del Open Data Movement, interesado en abrir su propia información para que los usuarios comunes pudieran trabajar con ella. Los datasets son publicados en un lenguaje llamado Resource Description Framework (RDF), el cual crea enlaces entre ellos para que la información pudiera ser reutilizada.

El objetivo de obtener un análisis cuantitativo y estructural de la Web of Linked Data es mejorar las búsquedas de datos. Teniendo un conocimiento profundo de la estructura y de las características de LOD, es posible mejorar el uso de dichos datos. Para trabajos futuros, las búsquedas de usuario podrían ser más rápidas y más precisas. En relación con este propósito nosotros nos aprovecharemos del uso del lenguaje de marcado Schema.org y del proyecto Linked Open Vocabularies (LOV).

Schema.org es un conjunto de etiquetas cuyo objetivo es que los Webmasters puedan marcar sus propias páginas Web con microdata. El microdata es usado

para ayudar a los motores de búsqueda y otras herramientas Web a entender mejor la información que estas contienen. LOV es un catálogo para registrar los vocabularios que usan los datasets de la Web of Linked Data. Su objetivo es proporcionar un acceso sencillo a dichos vocabularios.

En esta investigación, vamos a desarrollar un estudio que pudiera en un futuro ayudar mejorar las estrategias para buscar datos en la Web of Linked Data usando las fuentes mencionadas anteriormente con técnicas de "ontology matching". Estas técnicas tienen como objetivo mapear términos de diferentes fuentes de información para saber cuáles de ellos son comunes a ambas. En nuestro caso, primeros vamos a mapear Schema.org con LOV, y después LOV con la Web of Linked Data.

También se ha llevado a cabo un ARS de LOV. El objetivo de dicho análisis es obtener una idea cuantitativa y cualitativa de LOV. Sabiendo esto podemos concluir cosas como: cuales son los vocabularios más usados o si están especializados en algún campo o no. Estos pueden ser usados para filtrar datasets o reutilizar información.

Los hallazgos en este estudio muestran diferentes hechos. En el caso de la estructura de la Web of Linked Data, se concluye que es una estructura compacta y que las distancia entre nodos es baja. También se ha comprobado que cumple la teoría del bow-tie y que los datasets más importantes son WordNet 2.0 y DBpedia. En cuanto al análisis hecho en LOV, se obtienen las siguientes conclusiones: Los vocabularios no están especializados en un campo en concreto y no existe un dominio que sea más importante que el resto. También los vocabularios más importantes corresponden a estándares de la Web Semántica o son usados para modelar otros vocabularios como RDF, OWL o SKOS. Finalmente, con los mappings entre Schema.org, LOV y la Web of Linked Data, hemos desarrollado dos casos de uso de obtención de datos. El primero permite a los usuarios enriquecer páginas Web con información obtenida de datasets de LOD. El otro caso de uso consiste en ampliar ontologías con nuevas clases y propiedades procedentes de Schema.org. Un tercer caso de uso independiente que hemos mostrado consiste en obtener información de Google Scholar y agregarlo a fuentes de información que almacenan conocimiento científico como es el caso de VIVO y CERIF.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

CASRAI: Consortium Advancing Standards in Research Administration Information

CKAN: Comprehensive Knowledge Archive.

CRIS: Current Research Information Systems.

DBLP: Digital Bibliography & Library Project

FAO: Food and Agriculture Organization.

FCA: Formal Concept Analysis.

HTML: HyperText Markup Language.

HTTP: Hypertext Transfer Protocol.

LOD: Linked Open Data.

LOV: Linked Open Vocabularies.

LPHOM: Linear Program for Holistic Ontology Program.

MAP: Mean Average Precision.

METS: Metadata Encoding and Transmission Standard.

MODS: Metadata Object Description Schema.

N3: Notation 3.

NDCG: Normalized Discounted Cumulative Gain.

OAEI: Ontology Alignment Evaluation Initiative.

OWL: Web Ontology Language.

RDF: Resource Description Framework.

RDF-S: RDF Schema

RIF: Rule Interchange Format.

RSS: Really Simple Syndication.

SCC: Strongly Connected Component.

SKOS: Simple Knowledge Organization System.

SLDRM: Semantic Linked Data Retrieval Mode.

SNA: Social Network Analysis.

SPARQL: SPARQL Protocol and RDF Query Language.

SUMO: Suggested Upper Merged Ontolog.

SWPO: Semantic Web Portal Ontology.

SWR: Semantic Web retrieval.

SWRC: Semantic Web for Research Communities.

SWSE: Semantic Web Search Engine.

URI: Uniform Resource Identifier.

VOAF: Vocabulary of A Friend.

WWW: World Wide Web.

XSLT: eXtensible Stylesheet Language Transformation.

1. INTRODUCTION

Nowadays the World Wide Web (WWW) is a widespread. It consists of a set of documents or multimedia contents that are connected between them by links aimed to be readable by humans. In the structure, elements are identified by Uniform Resource Identifiers (URI) and information is retrieved using technologies as Hypertext Transfer Protocol (HTTP).

Like the Web of documents, it exists the Web of Linked Data a global data space of shared knowledge allowing data to be processed and understood by machines, (Heath & Bizer, 2011). This infrastructure arose when companies and organizations decided to publish their data and make them available for regular users. The problem became when the datasets started to be published using different languages and methods of access. To solve this problem Tim Berners Lee laid down the principles for publishing and interlinking data on the Web, (Bizer, Heath & Berners-Lee 2009).

The application of these principles led the Web of Linked Data to be seen as a graph. Datasets have been published exposing the information as triples allowing the information to be connected. The model of triples is based on the pattern “subject-predicate-object” in which a subject and an object are connected by their predicate. When the subject and the object of a triple belong to different datasets a link between them is created. This kind of connection allows the interchange of information between datasets and the creation of the graph structure commented before. In the graph, the datasets are the nodes and the links are the edges connecting them.

Similarly, as the early studies on the Web used to improve the design of search engines, here we make a study of the Web of Data. However, as the Web of Data has a different nature than the Web and revolves around the use of knowledge representations and schemas, we focus on three principal aspects: vocabularies, how these are used to annotate Web content, and the structure of Linked Data itself.

1.1 Motivation

1.1.1 Data retrieval strategies

One of the purposes of studying the structure of data-storage resources is improving retrieving data strategies or crawling. The literature in this field speaks most of the time of information retrieval, (Langville & Meyer, 2005), or data retrieval, (Gregory et al, 2017). The definition of information retrieval is mostly related with document searching. A good example of an information retrieval application could be Google. Instead, data retrieval is used for structured data with defined semantics and gives exact results or no results when querying the data. However, we will describe another classification based on our own point of view. Therefore, before studying the structure of the Web of Linked Data, we shall demonstrate that we can retrieve data from this source.

An example of applications with a data retrieval process is that of enriching other data resources for example connecting the classical Web with the Web of Linked Data. During the last years, Webmasters started to tag their Website by using microdata, which helps search engines, and other tools to understand better the information contained in them. One of these microdata vocabularies is called Schema.org and consists of a set of tags introduced by HTML5 created by Bing, Google and Yahoo! on June 2, 2011 by (Johnsen, 2012). Figure 1-1 shows how the use of microdata in Webpages has increased during the last year. In this Figure, the x axis corresponds to the last 16 months before September 2017 and the y axis the number of microdata tags in the top 1 million Websites by traffic.

So, there is an interest in connecting the tags contained in the Websites with the information of the Web of Linked Data datasets, we could retrieve information from it and aggregate it to Websites. As terms of the datasets are defined by vocabularies and we have a catalogue of these vocabularies in LOV, we could try to build a bridge between a set of Web tags like Schema.org and the Web of Linked Data by using LOV's vocabularies. We will use ontology matching techniques to achieve it.

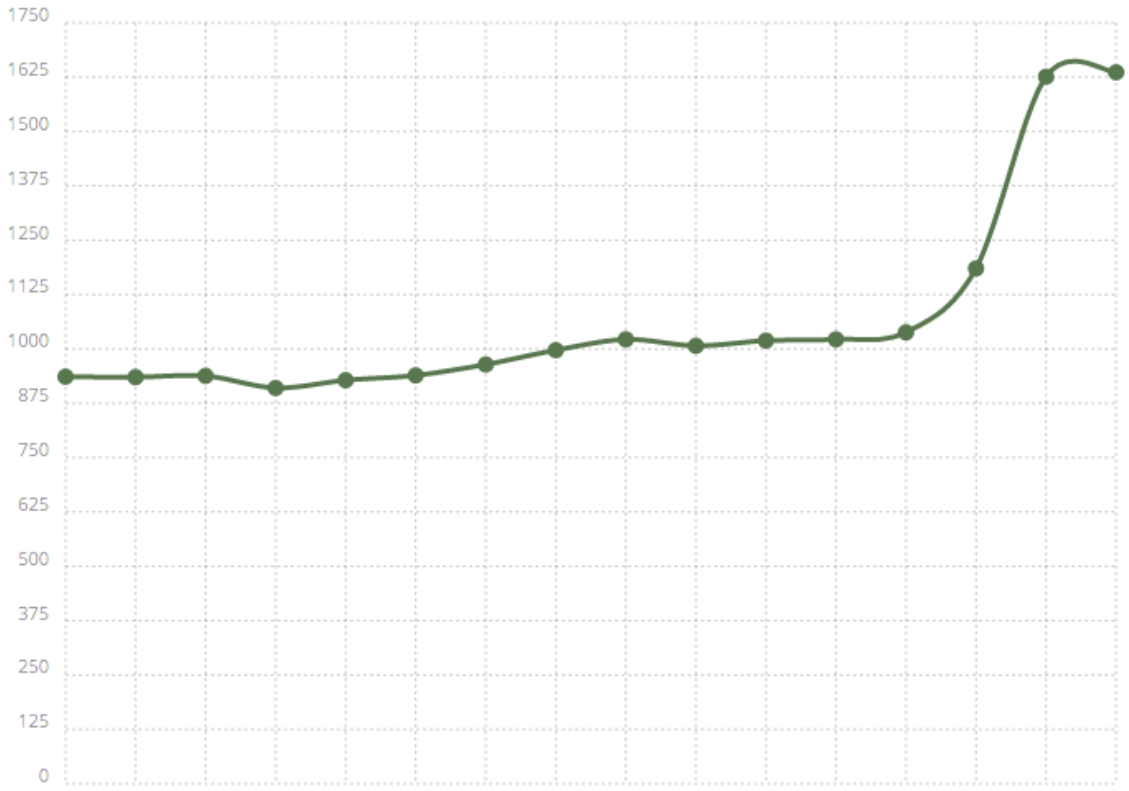

Figure 1-1: Use of microdata during 2017[1].

Ontology matching techniques find common terms between two ontologies or vocabularies. In our case, we will develop our own script that will work in two stages. The first stage will find which classes are shared by Schema.org and LOV and the second will do the same but concerning properties. The mappings will discover the terms that are equal string by string and those that are synonyms. Once we have this set of terms, we will study the impact of them in the Web of Linked Data datasets.

Finally, as a demonstration of data retrieval from one source to enrich another, we are performing two use cases. In the first one, we will retrieve information from LOD to enrich Websites using Schema.org tags. In the second one, we will extend a vocabulary from LOV using Schema.org terms.

In another experiment, we are making use of the RDF version of a dataset from the Web of Linked Data to design a data retrieval strategy. This case, focused in the field of scientific knowledge storage, will use three different sources OpenAGRIS, VIVO and Google Scholar. VIVO is an open source Semantic Web

[1] http://buildwith.com

application aimed to discover scientific knowledge based on an ontology and is part of the Web of Linked Data. OpenAGRIS is the RDF version of AGRIS, a catalogue of scientific publications in agriculture, also part of the Web of Linked Data. In this use case by searching papers from OpenAGRIS in Google Scholar, we will aggregate new information to a VIVO instance. Then the instance can be converted to CERIF, a standard of European Union written in XML presented in (Jörg, 2010), with a translator we have developed. The translator has been made setting manually mappings between the terms of both schemas. It allows to translate from VIVO to CERIF and vice versa.

1.1.2 The importance of the vocabularies

Vocabularies are an important element of datasets, allowing users to describe the information by using classes and properties. They classify the different terms and relationships of constraints using those terms. The complexity of the vocabulary is directly related with the amount of terms it contains.

The roles of vocabularies in Semantic Web can be defined in two ways. One of them is data integration, that occurs when terms in different datasets need a disambiguation. It also occurs in the discovery of new relationships. The other role is knowledge organizations, which happens when that libraries, museums or governments need to organize large collections of documents. Another important characteristic of vocabularies in general, is that they enable data to be interpreted by machines. The best practices in vocabulary usage, advise data providers to use popular vocabularies. In case of using their own vocabularies, its terms should be dereferenceable into W3C standards.

There are no mandatory vocabularies but as they are used some of them become more popular. There is an initiative called LOV, (Vandenbussche et al, 2015), for building a catalogue of these vocabularies. The objective of this project is to give access to the vocabularies, describe the relations between them and how they are linked with the Web of Linked Data.

As we have seen, both the Web of Liked Data and LOV are presented as graph structures containing massive data. By making a quantitative and structural analysis, users could benefit better from the information stored in it. For example, allow data providers to reuse vocabularies when creating new datasets or

obtaining the most complete datasets when users need to choose a particular one. The analysis cited before will be another part of our research.

1.1.3 Web of Linked Data structure

In this research, we are applying statistical methods and SNA to obtain a report of the structure of the Web of Linked Data. SNA is defined as the analysis of patterns of relationships among people, organizations, states and such social entities, (Jamali & Abolhassani, 2006). In other words, it lets us know how the different elements of the structure are related between them, providing us a general picture of it.

The application of SNA to improve data retrieval strategies can be found in several fields. Applied to the Web, (Brin & Page, 1998) presents the prototype of the search engine Google. In this paper, a quality ranking called PageRank, (Page et al, 1998), is applied to the link structure of the Web. It is also applied to databases like (San Martín & Gutierrez, 2006), where an improved data workflow is defined based on a special social network data model. As case study is used Digital Bibliography & Library Project[2] (DBLP), a scientific knowledge storage of computer science papers from journals and proceedings. Another field of application is social networks like Facebook or Twitter. In (Mincer & Niewiadomska-Szynkiewicz, 2012), SNA is applied to determine interpersonal connections, find principal actors and communities of people.

As said before SNA was first used in social science. For example, in (Adamic & Glance, 2005), the relations between political bloggers are studied trying to find the interaction degree between conservative and liberal communities. But we can find other fields in which applied these techniques are applied. In neuroscience, (Deco et al, 2013), SNA techniques are used to analyse resonance imaging biomarkers that could be used to explain different stages of Alzheimer. (Walther, 2015) in economics used it to better understand trading in developing countries from Africa. In the field of politics, (Koschade, 2006), tries to understand the communication and structure of terrorist cells, predicting its outcomes. Related with anthropology, a study to measure the impact of cultural organizations'

[2] http://dblp.uni-trier.de/

programs to regenerate a community is presented in (Oehler et al, 2007). Finally, in psychology, for example (Soares & Lopes, 2014) presents a social contagion model demonstrating that the central member of a network is dominant in psychological safety.

But the field that interests us is computer science, where we also have some SNA applications. In computer security, (Wang et al, 2009), SNA is used to build an intrusion detection system in mobile networks. In (Castillejo et al, 2012), is developed a recommendation system, a typical application in the field of artificial intelligence. In the field computer vision, (Renoust et al, 2015), applied it to a system of face detection in news videos. (Papadimitriou et al, 2009), used it in communication networks in particular in wireless sensor networks.

But if we have to point out a particular previous study, that will be (Broder et al, 2000). This paper makes a SNA of the structure of the Web of documents. Having a deep analysis of it, can help to design crawl strategies, analyze the behavior of Web algorithms or predict the evolution of the Web structure. If we take into account that the paper has been cited 3,549 times, regarding Google Scholar. That points out the importance of the study and the possibility to make a similar analysis of the Web of Linked Data. So, making this study is the one of the principal objectives of this research.

1.2 Objectives

The main objective during this work is to study the structure of the Web of Linked Data itself and its different components as a previous step to improve data retrieval in future works. Also, some use cases of data retrieval will be presented. The particular objectives are presented in what follows, and the specific steps or parts involved on each one as sub-objectives.

O1. Connect the Web of Linked Data with an independent data source. The aim of this objective is to take advantage of the information stored in the Web of Linked Data.

O1.1. Present a use case in which information from the Web of Linked Data is aggregated to an independent resource.

O1.2. Present a use case where a dataset of the Web of Linked Data is used to guide a data retrieval strategy.

O2. Make an analysis of the structure formed by the vocabularies used in the Web of Linked Data. This could also be used to perform better strategies to retrieve data from the Web of Linked Data or to optimise the use of vocabularies when developing new datasets.

O2.1. Make a quantitative report of the characteristics all the vocabularies have in common.

O2.2. Understand the structure formed by the relations of the different vocabularies.

O2.3. Report the usage of the different vocabularies in the datasets of the Web of Linked Data.

O3. Make a structural and quantitative analysis of the Web of Linked Data that could benefit users to improve data retrieval: make searches with more accuracy or retrieve data faster.

O3.1. Make an analysis of the overall structure. Obtaining general characteristics of the graph formed by the datasets of the Web of Linked Data.

O3.2. Make an analysis of the connectivity of the structure, so we can figure out which is the structure.

O3.3. Check if the structure of the Web of Linked Data accomplishes with the theory of the bow-tie to know how the datasets can be grouped by the way they are connected.

1.3 Structure of the document

Apart from this section, the rest of the document is structured as it follows:

In the second section, a background of the different fields involved in the research is proposed. This will let the reader to understand the basic terms of each field and how it has evolved. Also, will let them to know some previous researches addressing similar problems, so the input of the research could be proved.

In the third section, the studies carried out will be described for each the objectives listed in the first section. For each study, there will be an approach, information about the tools and resources used, how the data

has been collected, some reports about results and a discussion about them

In the fourth section, the last, the conclusions are exposed, relating them with the proposed objectives. Also, future lines of investigation will be set out.

2. BACKGROUND

In this section, the literature related with the research will be reviewed. This will be useful: first to understand the general concepts of what these research's lines are about and second to know the current state of the field and how this thesis contributes to this knowledge.

2.1 Internet, information resources and metadata.

What we call nowadays the Internet is defined as the Web of documents. It is a set of digital resources/objects connected between them by links. We define a digital object as a resource that is generated through some electronic medium and made available to a wide range of viewers both on-site and off-site via some electronic transferring machine or Internet, (Saye, 2001). Another definition is the one given by (Harvey & Thompson, 2010), as a compound object that must have these elements: the material, descriptive metadata, technical metadata, an activity/event log, representation information, and a unique identifier.

In order to store digital objects several technical infrastructures have been defined but two of them seems to be more used than the others; digital libraries and digital repositories. These definitions are based on visionary ideas from authors like Bush or Lickleder. Bush (1945) says: "The Encyclopaedia Britannica could be reduced to the volume of a matchbox. A library of a million volumes could be compressed into one end of a desk". Another idea is from Lickleder (1965): "We delimited the scope of the study, almost at the outset, to functions, classes of information, and domains of knowledge in which the items of basic interest are not the print or paper, and not the words and sentences themselves - but the facts, concepts, principles, and ideas that lie behind the visible and tangible aspects of documents".

Different definitions have been proposed for each concept. The most commonly accepted for digital library which can be consider the most popular one, are:

- Waters (1998): Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they

are readily and economically available for use by a defined community or set of communities.

- (Leiner, 1998): The digital library is the collection of services and the collection of information objects that support users in dealing with information objects available directly or indirectly via electronic/digital means.
- (Arms, 2000): A managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network.
- (Borgman, 2000): Digital libraries are a set of electronic and associated technical capabilities for creating, searching, and using information.
- (Smith, 2001): Digital library as an organized and focused collection of digital objects, including text, images, video and audio, with the methods of access and retrieval and for the selection, creation, organization, maintenance and sharing of collection.

Regarding the definitions for digital repositories, we have chosen the following five:

- (Crow, 2002): Digital repositories are commonly used for open access research outputs and regarded as an immediate and valuable complement to the existing scholarly publishing model.
- (Koutsomitropoulos et al, 2004): A Digital Repository is a collection of digital entities that are subject to the following three operations: insertion, deletion and retrieval.
- (Hayes, 2005): A digital repository is where digital content and assets are stored and can be searched and retrieved for later use.
- (Papparlardo et al, 2007): A digital repository is an online archive in which authors and academics can deposit their work, with the intention that it will be openly available in digital form.
- (Sharif & Uhlir, 2009): An online, searchable, web-accessible database containing works of research deposited by scholars. Its purpose is both increased access to scholarship and long-term preservation.

In the Internet-age, the term used to describe the data stored in digital libraries is metadata. (Bargmayer & Gilman, 2000) define metadata as “data about data”. In a deeper way, (Caplan, 2003) says that metadata is “structured information about an information resource of any media type or format”. In the digital library community, metadata is pointed as an important aid in discovery resource. Depending on its usage metadata are divided in different types. To avoid the different problems derived from this, metadata standards are created. Metadata standards provide structure and rules for the consistent provision of data. Depending on the field and type of data used. There exist primary standards like Metadata Object Description Schema (MODS), or Metadata Encoding and Transmission Standard (METS). MODS is a schema to be used with a bibliographic element set, (Guenther, 2004). Regarding (Cundiff, 2004), METS is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library. In a more general way we have Dublin Core or Schema.org. The Dublin Core Schema is a set of terms creating a vocabulary that can be used to describe both web and physical resources. Schema.org has preciously ben describe as a vocabulary that help search engines to improve their results. Also exists SKOS, which are specifications and standards to support the use of thesauri, classification schemes, subject heading lists and taxonomies within the framework of the Semantic Web.

2.2 The Semantic Web

The previous subsection introduces the term “Semantic Web”, presented in (Berners-Lee et al, 2001). As the Internet became more famous, the amount of online resources grew and turned more complex so obtaining the accurate information became more difficult for users. If users had problems to retrieve information, in the case of automated processes in which computers must understand the data, it was even more difficult. In this sense arose the need of adding structured and enriched content using semantic information so computers could understand it and applications could process it automatically. So, Semantic Web is understood as an extension of the traditional Web, whose definition in the previous paper from 2001 is “The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation”.

In the following Figure, taken from (Fensel et al, 2011)., the architecture of the Semantic Web is shown. It is divide in different layers, each with its own components. Starting from the bottom of the architecture, we find two layers:

- URI+Unicode and XML. The functions of these layers are two: mechanisms to univocally identify the concepts and which is the format of the messages.
- The next two layers, formed by SPARQL Protocol and RDF Query Language (SPARQL), RDF Schema (RDF-S), OWL and Rule Interchange Format (RIF), accomplishes with one of the aims of the Semantic Web: the development of a worldwide knowledge base.
- The Logic layer is used to define the logic that will be applied over the knowledge define in the previous layer.
- The last two layers, Proof and Trust, ensure that the information given by the other layers is valid and should be believed by the interchanging agents.

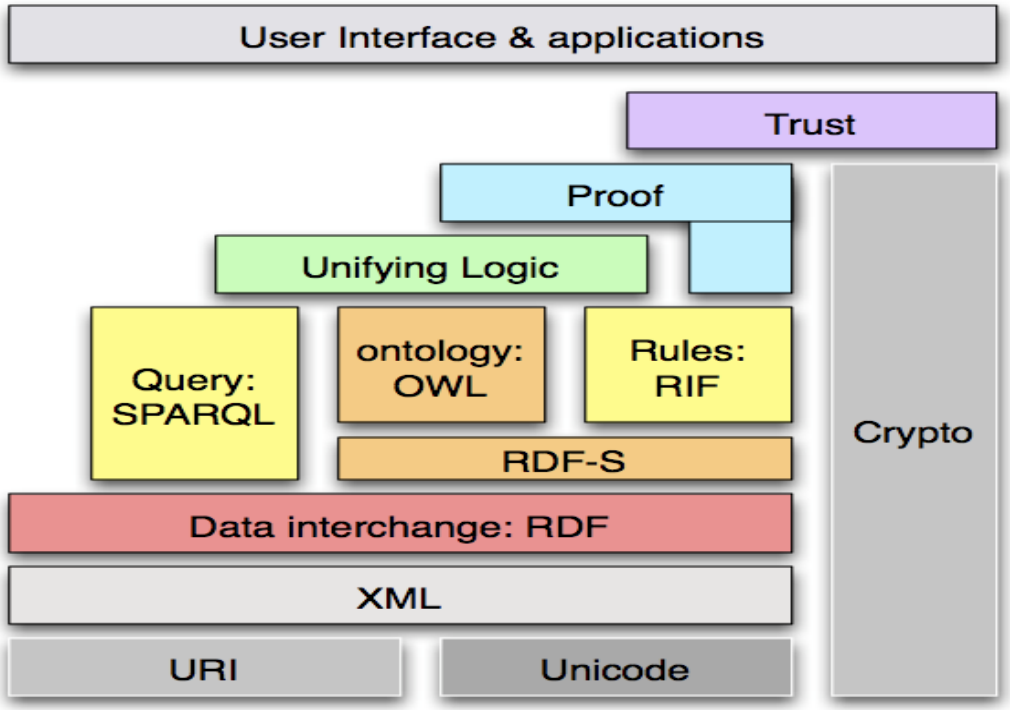

Figure 2-1: Semantic Web architecture. (Fensel et al, 2011).

In turn, each layer is structured around a set of components. They are:

- **URI**[3] is the way to univocally identify a resource.
- **Unicode**[4], is a standard for encoding and representing texts and characters.
- **XML**[5], eXtensible Markup Language, is a syntax that gives meaning to the content of a Website by using tags. As HyperText Markup Language[6] (HTML) is used to differentiate the visual parts of a Web, XML allows describing the information contained in it.
- **RDF**[7], Resource Description Framework, is a language aimed to represent and exchange data in the Web.
- **RDF-Schema**[8] is an extension of RDF allowing the representation of vocabularies.
- **SPARQL**[9] is the query language for RDF.
- **OWL**[10], is the language used to encoding and exchanging vocabularies in the Semantic Web.
- **RIF**[11] was developed for to interchange rules in rules-based systems of the Semantic Web.

But the most important components are, undoubtedly, the ontologies. An ontology is defined by (Gruber, 1993) as “the specification of a conceptualization”. Thus, with an ontology we describe a set of concepts and their relations, so the knowledge can be shared and reused. An ontology is formed by various components, (Slimani, 2015):

[3] https://www.w3.org/wiki/UriSchemes

[4] http://www.unicode.org/versions/Unicode6.1.0/

[5] https://www.w3.org/TR/2006/REC-xml11-20060816/

[6] https://www.w3.org/TR/html51/

[7] https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

[8] https://www.w3.org/TR/rdf-schema/

[9] https://www.w3.org/TR/sparql11-query/

[10] https://www.w3.org/TR/owl-features/

[11] https://www.w3.org/2001/sw/wiki/RIF

- **Concepts**: a group of individuals that shares common characteristics used in a wide sense.
- **Relations**: describe the means in which individuals (instances or particulars) are related.
- **Functions**: are particular types of relations, where the n^{th} element of the relationship is distinctive for the n-1 preceding elements.
- **Axioms**: represent assertions formulated in a logical form that together comprise the core knowledge that the ontology describes in its domain of application.
- **Instances**: are individuals that models particular objects (people, proteins, machines) and represents the base components of an ontology.

It is important to make a difference between ontologies and vocabularies, as they seem to be the same., An ontology is mostly used for a more complex set of terms with interrelationships or axioms and vocabulary is used when formalism is loose.

The W3C decided in 2003 to adopt OWL as a recommendation to represent ontologies. OWL has more expressivity power than other languages like XML or RDF. RDF-S can be used to represent simple ontologies but if we want to define more complex ontologies we need to use OWL as it allows expressing logic in the Semantic Web. It uses predictive logic to express constraints between classes, entities and properties. In fact, OWL has three different expressive languages each designed for a different purpose:

1. **OWL Lite** supports those users primarily needing a classification hierarchy and simple constraints. For example, while it supports cardinality constraints, it only permits cardinality values of 0 or 1.
2. **OWL DL** supports those users who want the maximum expressiveness while retaining computational completeness (all conclusions are guaranteed to be computable) and decidability (all computations will finish in finite time).
3. **OWL Full** is meant for users who want maximum expressiveness and the syntactic freedom of RDF with no computational guarantees.

2.3 The Web of Linked Data

When Semantic Web technologies started to become more popular, different companies and organizations started to publish their datasets. The problem was that the data was published in various formats and had different mechanisms of access. To solve that problem Tim Berners Lee laid down the four principles to publish interlinking structured data on the Web.

These four principles, presented in (Bizer, Heath & Berners-Lee, 2009), are:

- URIs should be used to identify things.
- More specifically, HTTP URIs should be used so that people can look up things.
- If someone looks up a URI, useful information should be provided using standards (RDF, SPARQL).
- Further semantic links should point to other URIs so that people can discover more things.

The first principle is about using URIs to identify a resource. It aims the resources to have a unique identifier, so other resources could reference them allowing to reuse the information. The second one, allows to use the HTTP protocol to query information. By the third principle if a user searches and accesses a resource with an URI, the provided information must be given in RDF format easily readable. The last principle let the users to connect resources from different places sites. By applying this, datasets are not being isolated, and the information is being reuse.

The data will be published using the four principles presented above and expressed in RDF by using triples. Triple is the basic concept of Linked Data. It consists of publishing data using the structure of subject, predicate and object. Subjects are resources represented by an URI. Objects could be other resources or particular values. Finally, predicates are also represented by URIs and are the way to know how subjects and objects are related. In the following Figure, it can be see a graphical representation of a triple and its corresponding code in RDF. In that example “Person” is the subject, “name” the predicate and “Henry” the object.

```
@prefix ex: <http://example.org> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

ex:Person foaf:name "Henry"
```

Figure 2-2: Graphical representation of triples and its RDF code.

By adding triples sharing nodes, we obtain a set of connected triples that we called RDF graph. This structure can be seen as a directed graph where the subject and object are the nodes and the predicates are the directed edges. In the following Figure, there is an example of an RDF graph. Here we can see an RDF graph formed by three triples: first one is "Person-name-Henry", second is "Henry-worksin-Alcalá University" and the last one "University-name-Alcalá University". With this graph, we are saying that there is a person called Henry that works in Alcalá University, which is a University.

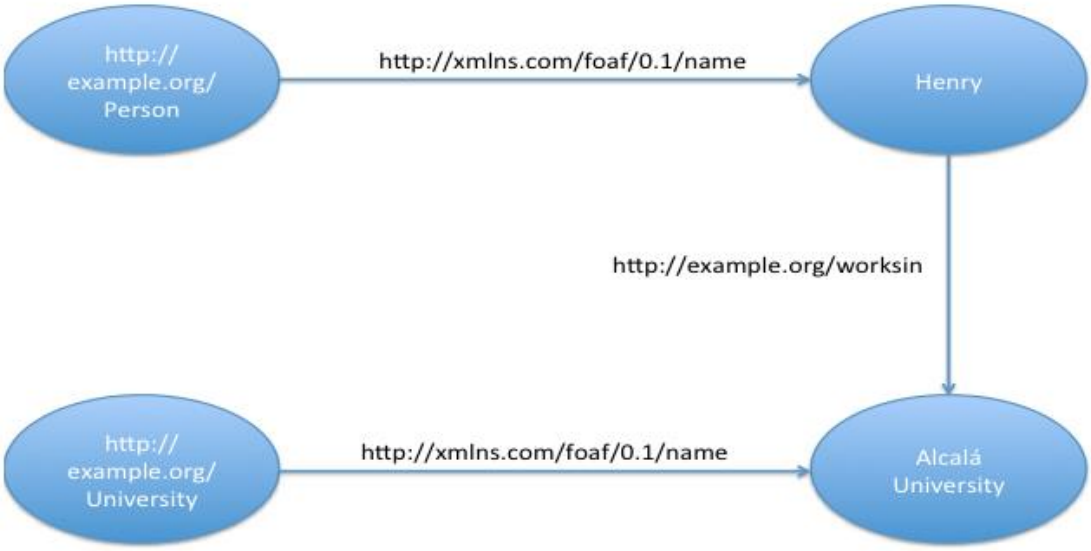

Figure 2-3: Graphical representation of an RDF graph.

The RDF graphs can be serialized in three languages:

- **RDF/XML**[12] that is a syntax created to express an RDF graph in XML. It has the problem that is not easily readable by humans and due to his expressiveness cannot represent some graphs.
- **N3**[13] (Notation 3), which was created based in human-readability. We can consider N3 to be much readable and compact than XML.
- **Turtle**[14] and **N-Triples**, which are subsets of N3 allowing to represent triples in an easier way.

We now have to store the information in an RDF graph, but we do not know how to have access to it. For that purpose, we have SPARQL that allows to obtain particular parts of a graph by using a similar syntax to the database languages. By creating queries, the user can retrieve or manipulate the data stored in the graph. A basic example of a query can be seen in the following Figure. In the query, we have five different parts. The first one is for the prefixes we are going to use to retrieve the information. The second part is for the dataset/graph in which the information is stored. The next one is the variable we want to obtain, in this case the subject of a triple. Then, we have a triple that has to match with the information we need to get; here we only need those which have the predicate schema:name. Finally, we have the modifiers that we use to order or limit the results.

[12] https://www.w3.org/TR/rdf-syntax-grammar/

[13] https://www.w3.org/TeamSubmission/n3/

[14] https://www.w3.org/TR/turtle/

Figure 2-4: SPARQL query and its different parts.

By publishing collections of RDF graphs interconnected between them, we get RDF datasets. This RDF datasets need to follow the four principles of Linked Data explained before. Then when a dataset needs to reuse a term already published by another dataset, an RDF link is created. By connecting different datasets, we obtained what is called the Web of Linked Data. This Web is similar to the classical Web of Documents but instead of Webpages, we have datasets and instead of links between Webpages, we have RDF links.

After some years a few governments and organization started to make their data publicly available for any user. This aims to reuse and enrich the different datasets. With the aim that people started to publish their dataset and increase the number of Linked Open Data, Tim Berners Lee proposed a 5-star-model in 2010[15]. The following table shows a summary of the proposal.

[15] https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data

Star	Description
1 star	Data is available on the Web (whatever format), but with an open license.
2 stars	Data is available as machine-readable structured data (e.g., Microsoft Excel instead of a scanned image of a table).
3 stars	Data is available as (2) but in a non-proprietary format (e.g., CSV instead of Excel).
4 stars	Data is available according to all the above, plus the use of open standards from the W3C (RDF and SPARQL) to identify things, so that people can link to it.
5 stars	Data is available according to all the above, plus outgoing links to other people's data to provide context.

Table 2.1: 5 stars model for Linked Open Data.

As the publication of more datasets in the Web of Linked Data occurred, the need of having an idea of the structure formed by them emerged. The project aimed to study this is the LOD Cloud, Linked Open Data Cloud, that is a catalogue of the datasets published as Linked Data. In 2007, it had only twelve datasets, but by 26th of January 2017 it has 1,146 divided in nine areas. We can find datasets of cross-domain, publications, geography, social network, government, user generated, life sciences, linguistics and media. If we have to highlight a dataset, this is the central one called DBpedia, (Auer et al, 2007), which is an RDF version of Wikipedia. By following the different links between the datasets, we reach the information reused by another dataset. The following Figure shows the last state of the LOD Cloud.

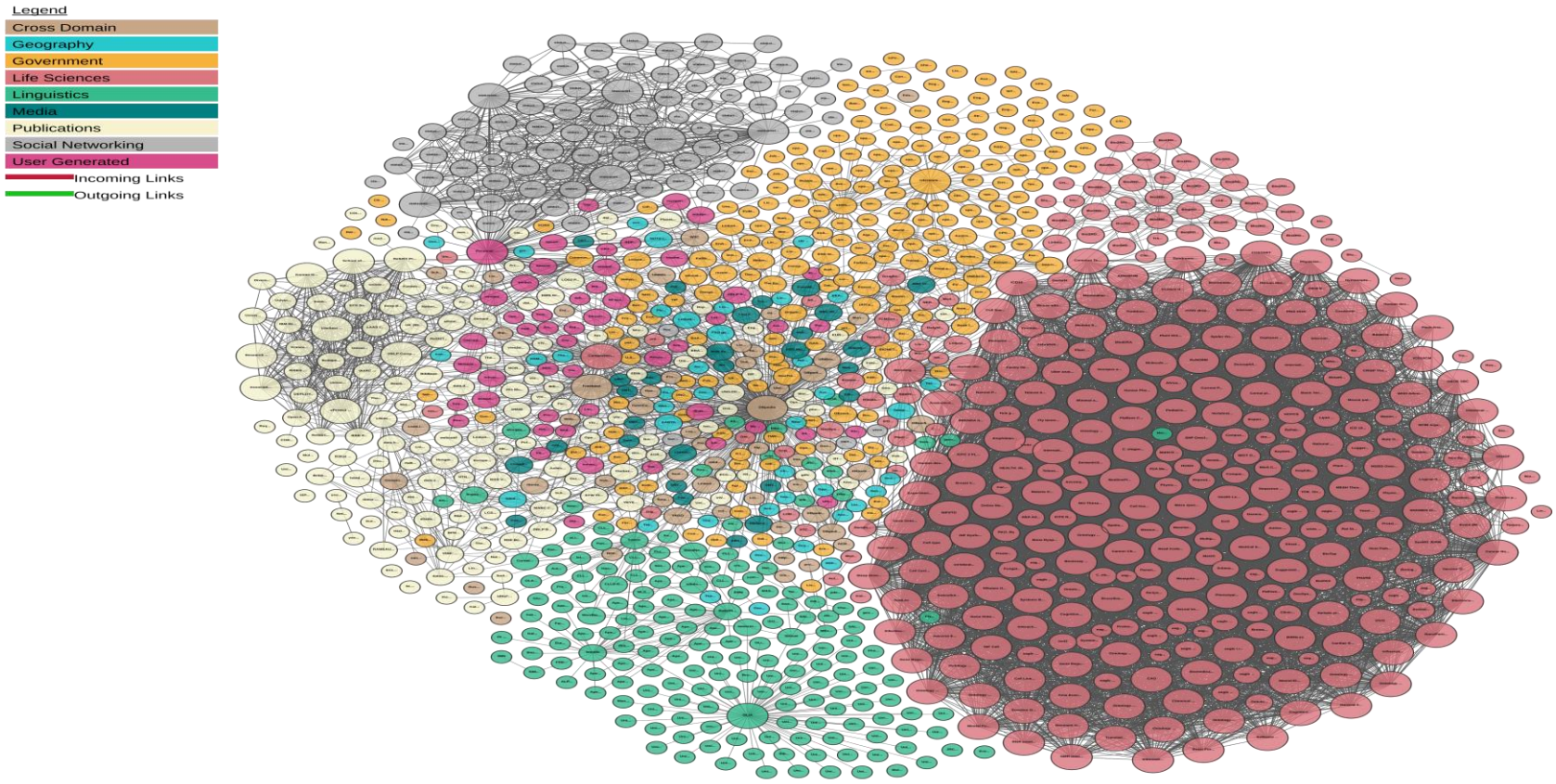

Figure 2-5: Linked Open Data diagram[16]

[16] http://lod-cloud.net/

2.4 Studies about the Web of Linked Data.

As there is a lot of information stored in the Web of Linked Data there is an importance in knowing its structure. If we have an idea about its actual state and how the information is distributed, better data retrieval strategies can be design or can be understood how datasets are related between them. We can analyse this kind of structures by using network and graph theories as SNA.

There are several papers trying to analyse the structure, metrics and usage of the Web of Linked Data. The following papers only work with little sets of LOD. A metric like semantic distance is calculated and applied in resource recommendations using in some cases DBpedia, (Passant, 2010). Then, in (Hoser et al, 2010), SNA is applied to two different ontologies: Suggested Upper Merged Ontology[17] (SUMO) used to describe all the concepts for merging ontologies of different domains, and Semantic Web for Research Communities[18] (SWRC) ontology with vocabulary to express research knowledge and its relations. The analysis goes from degree centrality or betweenness centrality to eigenvector centrality. Characteristics like size and if the distribution and complexity follow a power law are calculated in (Ding & Finin, 2006). The big difference here is that the snapshot of the Semantic Web is not the LOD Cloud as they harvest their own data having an own vision of the Semantic Web. Finally, in (Cheng & Qu, 2008) are analysed metrics like degree, reachability or connectivity of a dataset of 401 million triples.

In the following papers the SNA goes deeper or is applied to a general structure of the Web of Linked Data. In (Hausenblas et al, 2008), using a set of 34 linked datasets that could represent the Web of Linked Data. It is analysed: the size and accessibility of the datasets, also how the different datasets are internally and externally interlinked. An analysis of the Linked Data Cloud state is made on February 2009, in (Rodriguez, 2009). It reports general metrics like number of vertices and edges, if it is weakly or strongly connected, diameter and average path of length. It also makes a structural analysis where independent datasets

[17] http://www.adampease.org/OP/

[18] http://ontoware.org/swrc/

and domains are analysed. LODStats[19] is presented in (Auer et al, 2012), a framework to obtain statistics from datasets stored in Comprehensive Knowledge Archive[20] (CKAN). These datasets are serialized in RDF or are accessible via SPARQL endpoint. Analysis like the size of datasets and its evolution can be found. In (Dividino et al, 2014) some metrics of different snapshot of the Linked Data cloud are analysed in order to measure how it evolves during the time. Finally, in (Schmachtenberg et al, 2014), is presented the biggest crawl of the Linked Data cloud at the moment of this work. In this case only the linkage relationships between datasets is analysed.

2.5 Social Networks Analysis.

In (Freeman, 2004) the framework of SNA is developed. Here SNA is divided into four principal features: structural intuition, systematic empirical data, graphic representation and mathematical or computational models.

The first works in structural intuition were published by Henry Comte between 1830 and 1843. Other publications come from Henry Maine (1861/1931) or Ferdinand Tönnies (1855/1936). Talking about systematic empirical data, the first publication, (Huber & Bonnet, 1792), is a description of honeybees' behaviour, in which bumblebees demonstrate its dominance with respect to one another.

The first systematic data collection based on humans was ethnography of the Iroquois, (Morgan, 1851). Also, (Morgan, 1851) published the first graphical representation of relational data with a system of descent in the ancient Rome. (McFarlane, 1883) constructed a visual representation of various degrees of kinship. (Hobson, 1884) shows a picture to demonstrate how a small set of large corporations could control other firms by using interlocking directorates. Finally, there are the mathematical and computational models, where we can find the graph theory in (Euler, 1736).

SNA started to be applied during 1960s and 1970s by sociologist. Then statisticians, mathematicians and computer scientists were interested in the discipline leading to a fast development and application in several fields like

[19] http://stats.lod2.eu/

[20] https://datahub.io/

economics, marketing or industrial engineering, (Scott, 2000). Figure 2-7 shows how the usage of "social network analysis" in papers has increased and is still increasing through the years.

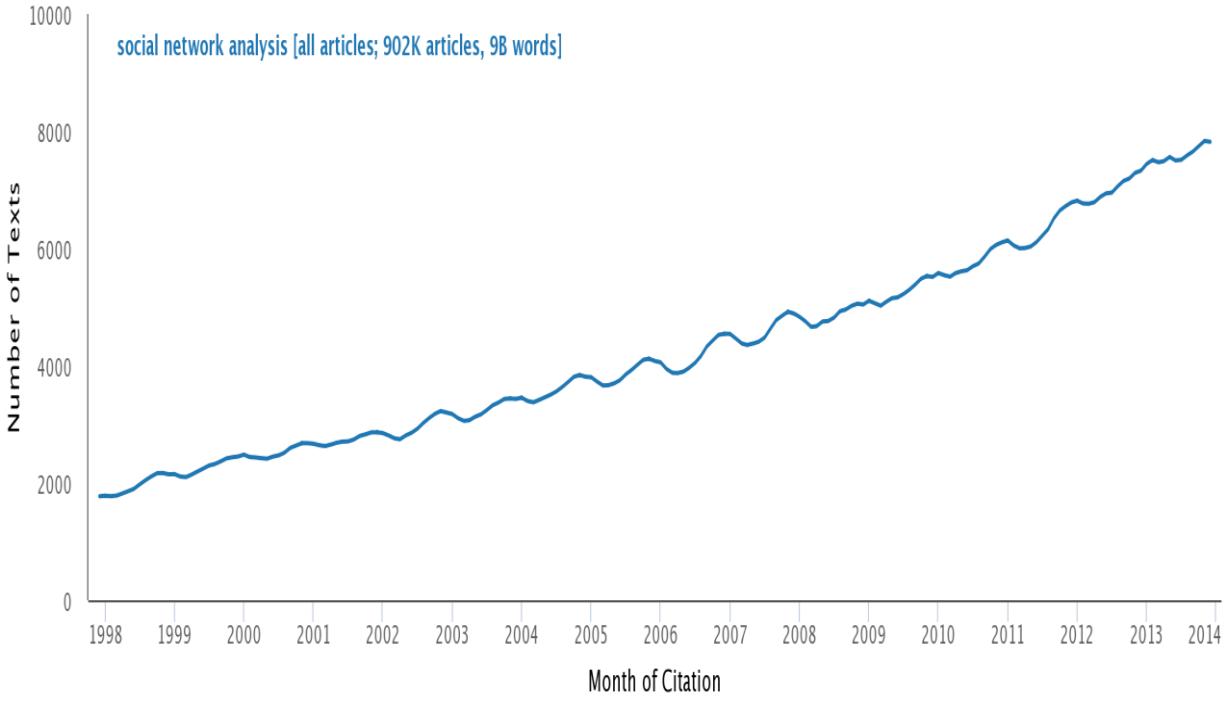

Figure 2-6: Use of the keyword "social network analysis" in papers[21].

The publication of books like (Wasserman & Faust, 1994) and the aforementioned (Scott, 2000) where a deep analysis of these techniques is made, started to increase its usage. Also, the development of SNA tools and packages that could analyse big amount of data helped. For example, EgoNet[22] used to analysed egocentric networks. In (Gansner & North, 2000), Graphviz[23] an open source graph visualization tool is presented.

There exists Python modules like graph-tool and NetworkX. The first one is used for manipulation and statistical analysis of graphs, (Peixoto, 2014). The second one is used for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks, (Hagberg, 2008). NodeXL[24] provides network

[21] http://arxiv.culturomics.org/

[22] https://sourceforge.net/projects/egonet/

[23] http://graphviz.org/

[24] http://nodexl.codeplex.com/

graphs exploration, metrics and sentiment analysis, (Smith et al, 2010). Finally, we have two of the most famous applications Pajek[25] and Gephi[26]. Pajek, (Batagelj & Mrvar, 2002), is a package for analysis and visualization of large networks. Gephi, (Bastian et al, 2009), is a visualization and exploration software for all kinds of graphs and networks.

The Web of Linked Data can be modelled using the same techniques used for Social Networks. In (Klez & Hodic, 2014), a Social Network is defined as a chain of individuals and their personal connections. In (Khoshnood, 2012) it is defined as a social structure of people having relationship based on casual interests e.g. friendship and honesty. Another definition is that given by (Srivastav & Nath, 2015), where a social network is considered to be a social structure made up of individuals (or organizations) called "nodes", which are tied (connected) by one or more specific types of interdependency, such as friendship, kinship, common interest, financial exchange, likings or disliking, or relationships of beliefs, knowledge or prestige.

A social network, in mathematical context, (Schaeffer, 2007), can be formulated as a graph G, which is a pair of sets $G = (V, E)$. V being the set of vertices (the number of vertices $n = |V|$ is the order of the graph) and E containing the edges of the graph.

We can define Social Network Analysis as mapping and measuring of relationships and flows between people, groups, organizations, animals, computers or other information/knowledge processing entities, Jamali and Abolhassani (2006). For Mincer and Niewiadomska-Szynkiewicza (2012) it is a group of graph theory based techniques that can be used to retrieve meaningful knowledge from networks formed by various actors. This last definition is the one that fits most with our thesis as we are interested in have a global picture of the whole structure of the Web of Linked Data but also at level nodes.

If we talk about the history of SNA, most of the literature agrees that the first researches of SNA were realized by psychiatrist Jacob Moreno (1889-1974). In (Moreno, 1932), he applied it to a community of prisoners. Later in (Moreno,

[25] http://mrvar.fdv.uni-lj.si/pajek/

[26] https://gephi.org/

1934), the case of study was a girls' reformatory. In 1938 alongside with psychologist Helen Jennings, they presented a technique in social configurations based on statistical treatment, (Moreno & Jennings, 1938). These researchers named as sociometric analysts and are considered to use and develop the graph theory. Simultaneously, a group of researchers in Harvard Business School, leaded by W. Lloyd Warner (1898-1970) and Eltan Mayo (1880-1949) focused on the industrial productivity. In (Warner & Lunt, 1941) a new concept of clique was introduced. (Mayo, 1945) started with the large usage of sociograms. Another trend of SNA emerged with Kurt Lewin (1890-1947), as a psychologist he applied the field in social psychology, (Lewin & Lippitt 1938).

First applied to social science, SNA has evolved being used in several fields. In medicine, (Tsalatsanis et al, 2011), used SNA to study the impact of interactions between randomized control trials on treatment success. In (Novielli & Marzak, 2013), in the field of software engineer, SNA is used to discover the relationship between developers in distributed teams. Also, SNA has been used in economy, for example (Koochakzadeh et al, 2012), to develop a recommender for non-expert investors. The size of the communities where SNA is applied also varies from little communities, like classrooms in (Grunspan et al, 2014), to nations in (Valente et al, 2015). We also have to review the utility of SNA, regarding (Pattison, 1993), two of them are the most used. First one is used to explain individual behaviour, the second tries to understand the social behaviour of a group.

One of the techniques applied to SNA is graph theory, where social networks are treated as graphs as we have defined before. For the purpose of this study the Web of Linked Data has to be considered as a directed graph. This is defined by (Kannan et al, 2008).

Definition 1: A directed graph G is a unique-path graph with respect to a source vertex s if there is at most one simple path from s to any vertex $v \in V(G)$.

Directed graphs are used when there is an interest on how the information is flowing between nodes. For example, in the case of paper citations, citing a paper has to be differentiate of being cited by a paper.

Taking that into account, the Web of Linked Data can be formal defined in (Passant, 2010) as:

Definition 2: A dataset following the Linked Data principles is a graph G defines as $G = (R,L,I)$ where

- $R = \{r_1,r_2,...,r_n\}$ is a set of resources - identified by their URI
- $L = \{l_1,l_2,...,l_n\}$ is a set of typed links - identified by their URI
- $I = \{i_1, i_2, ..., i_n\}$ is a set of instances of these links between resources, such as $i_i = \langle l_j,r_a,r_b \rangle$.

Scaling to the Web, the Linking Open Data cloud is then defined as the union of all the graphs Gi that are published (and interlinked) on the Web, i.e. $LOD = \cup i\ Gi$.

When applying graph theory and SNA, the user tries to obtain some metrics. There are general metrics like number of nodes and edges, type of connectivity, diameter, density or centrality. Then, regarding the nodes the degree distribution (in-degree and out-degree) can be calculated. Finally, the connectivity based on the bow-tie graph theory is an important metric to take into account. Definitions related with that are found following.

Definition 3: A connected component, (Feng et al, 2016), is a maximal subgraph of a graph in which any two vertices are connected to each other by a path.

Connected components let us to discover local communities in networks.

Definition 4: A directed graph is strongly connected if there is a path between all the pairs of nodes. If we have a maximal strongly connected subgraph, we can consider it a Strongly Connected Component (SCC) of a graph. It can be calculated by using Tarjan's algorithm, (Tarjan, 1972), with Nuutila's modifications, (Nuutila & Soisalon-Soininen, 1999).

Definition 5: A weakly connected graph is when avoiding the directions of the edge it becomes a strongly connected graph.

Definition 6: Effective diameter or eccentricity as, proposed in (Tauro et al, 2001), is the minimum number of hops in which some fraction (say, 90%) of all connected pairs of nodes can reach each other.

Diameter gives us an idea of how easy the information can be expanded over all the nodes of a network. A low diameter means that is easier to reach all the nodes of the graph starting from a particular one.

Definition 7: Density the proportion of edges compared to the maximum edges of the graph if it were complete.

Density can be used to know how fast the information is spreading among the network.

Definition 8: Reachability measures the number of nodes to go from one to another, no matter how many you have to pass through.

This measure tells us if a node is more isolated than others drawing possible divisions in the network.

Definition 9: Degree centrality as, proposed in (Opsahl, Agneessens & Skvoretz, 2010), taking into account that the degree of a node is its number of connections, was computed as the number of ties or neighbours of a node.

Talking about undirected graphs, a node with a lot of edges has more possibilities to obtain the information that is flowing in the network. That means that the node will be less dependent to the rest of the network. In directed graphs we have to differentiate between edges reaching a node and those leaving it, we call it in-degree and out-degree respectively. In-degree has to be interpreted as popularity and out-degree as influence.

Definition 10: Closeness was the inverse of the sum of all shortest paths to others or the smallest number of ties to go through to reach all others individually.

The closeness centrality emphasizes in the distance of a node to reach the others. These nodes having a lower closeness centrality are considered a reference point in the network, so spreading information starting from this point will cost less.

Definition 11: Betweenness centrality, introduced by (Freeman, 1977), is a way to measure how a node can control the relations between other nodes in a social network.

Betweenness centrality of a certain node (its actor centrality) will be given by the proportion of times it is between other nodes for sending information and the number of falls in pathways between other nodes.

Definition 12: Eigenvector centrality. It based on the idea that if a node influenced another node and this one is influencing other, the nodes at the end of the chain will be highly influential.

It is used when there is an interest in ranking the nodes of a network in terms of popularity. Taking into account that a node is popular not only if it has a lot of

friends that could be reached in one step. These friends also have to be popular and have to be connected to a lot of nodes. It is similar to how Google ranks the Websites.

2.6 Ontology Matching.

As ontologies started to become more popular, they started to be developed by different actors. At that time, it arose the problem of heterogeneity, when a new ontology needs to be integrate in a system. It could happen that different ontologies describing the same item would use different terms, so when exchanging information, the system would understand that they were referring to different things. To solve that problem, they started to use ontology alignment. For example, in a traffic system, one ontology could use “speed” to determine the speed limit in a road and another one “velocity”. Ontology alignment sets mapping between terms that are semantically equivalent, so the system could understand that they refer to the same, solving the problem. The process used to find these mappings is called ontology matching. In (Ehrig & Euzenat, 2005), a formal definition of ontology alignment can be found.

Definition 13: Given two ontologies O and O', an alignment between O and O' is a set of correspondences: $<e,e',r,n>$ with $e \in O$ and $e\,0 \in O'$ being the two matched entities, r being a relationship holding between e and e', and n expressing the level of confidence $[0..1]$ in this correspondence.

As we said before the problem is related with heterogeneity. There are several types of heterogeneity not only the one named above. Due to that, different classifications of heterogeneity have been exposed during the last years. But if we read works related to the field, most of them talk about the classification presented in (Euzanet & Shvaiko, 2007):

- **Syntactic heterogeneity**: due to the different ontology languages that could be used. Also, when using different languages in the used vocabularies, for example Spanish vs German. Another example occurs when the level of formality is different. To solve the problem equivalences between the languages or between tags are set.
- **Terminological heterogeneity**: this happens when the two terms have the same meaning: when talking in different fields or synonyms. For

example, article is used in journalism and paper in research or firm and house, which are synonyms.

- **Conceptual heterogeneity**: it is when there are differences trying to model the same domain. This means that different axioms are used to define concepts or just the use of different concepts. Depending on the reason a classification is given:
    - Difference in coverage: it happens when two ontologies describe different parts of the domain with the same level of detail and perspective. For example, a factory of beers and a factory of screens. Both are factories and have same characteristics but the process to manufacture the products is different.
    - Difference in granularity: it happens when two ontologies describe the same part of a domain but with different levels of detail and same perspective. For example, an ontology about sports could describe the main characteristics of a soccer team and another one also its players.
    - Difference in perspective: it happens when two ontologies describe the same part of the domain, with the same level of detail and different perspective. For example, a map in agronomy can be defined for crops or to measure the levels of irrigation.
- **Semiotic heterogeneity**: this case depends on people's view. Due to its nature is really difficult to detect by computers and also to solve.

To solve heterogeneity problems, matching techniques have to be applied. Based on these different classifications about the techniques can be found, they depend on the focus. A basic classification, which covers the minimum requisites, is in (Giunchiglia & Shvaiko, 2003). In this paper, the classification is based on how matching elements are computed, having syntactic and semantic matchings. The first one measures how similar are two terms, for example "car" and "car" are exactly the same and "phone" and "telephone" are almost the same. The other classification occurs when two terms that are totally different words have the same meaning, as synonyms. But if we have to take a classification, which is considered the most complete, this is (Euzenat & Shvaiko, 2013); a figure depicting it is shown below.

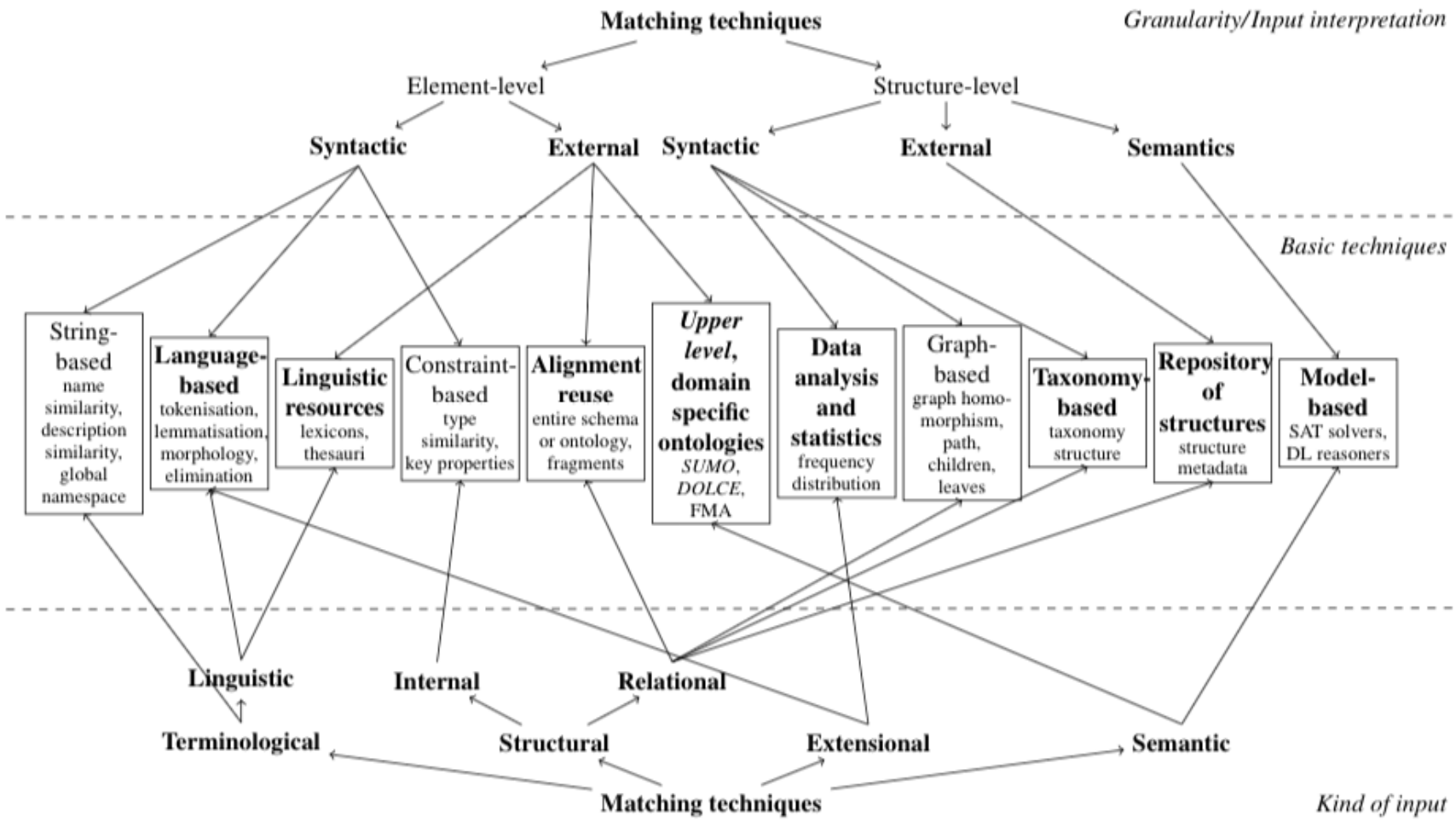

Figure 2-7: Matching techniques classification, (Euzenat & Shvaiko, 2013).

This classification can be read in two ways. The top-down view has a first level based on the “Granularity” and a second level based on “Input Interpretation”. The “Granularity” level can be divided into "Element-level where the entities of the ontology are considered as unique elements and not as part of the whole ontology and Structure-level where the entities are analysed with respect to the whole ontology. The Input Interpretation is divided into: syntactic, external and semantic. Syntactic techniques interpret the input by itself using some algorithms. External techniques interpret the input with the help of external resources of the same field and common knowledge. Finally, semantic techniques use some formal semantics.

Considering the different classifications presented above, it arises a set of techniques that are being explained following:

- String-based: usually used to match classes and properties from ontologies. It is based on comparing terms string by string. So, if to strings are equal letter by letter then they refer to the same term.
- Language-based: these techniques are based on the idea of Natural Language Processing. Here, terms are not only considered as simple strings. Their roles in a sentence or the used language are important. Technique like tokenization or lemmatization or the use of thesauri and dictionaries are included here.
- Linguistic resources: we refer to them when we combine linguistic relations like synonyms and antonyms with the use, for example, of thesauri in a particular domain
- Constraint-based: to calculate similarities, this takes into account internal characteristics of the entities as range or cardinality in properties or type to specify the instances.
- Alignment reuse: it is based on the idea of reusing an alignment of two or various ontologies to align one of these to another ontology.
- Upper level, domain specific ontologies: taking into account that these kinds of ontologies cover a set of general concepts. They can be used as a starting point to create a domain specific ontology just reusing the upper concepts that they will share.
- Data analysis and statistics: these techniques take a representative set of instances of the population trying to find subsets with common characteristics or calculating the distances between them.

- Graph-based: in this classification, we find algorithms that are working with ontologies as they were labelled graphs. The main idea behind these techniques is that the neighbours of similar nodes from two different ontologies are somehow similar.
- Taxonomy-based: are also graph algorithms but taking advantages of the "specialisation" relations between nodes. Similar to the idea above, nodes related by "is-a" property have nodes that are somehow similar.
- Repository of structures: it creates a repository of ontologies and their fragments, storing their pairwise similarity. This information, alignments are not created, is used when adding a new ontology or fragment of an ontology. If the similarity with the data stored in the repository is high, then is it worthy to do a more exhaustive analysis.
- Model-based: also called semantic grounded, it uses the semantic interpretation to obtain alignments. If two terms are the same they will have the same interpretation.

Now that we have a classification of the different ontology alignment techniques, we can do a little survey of ontology alignment tools. Taking into account that it is a hot topic, every year new tools are developed. A good way to know the state of the field is to pay attention to the Ontology Alignment Evaluation Initiative[27] (OAEI). It is a yearly event where different ontology alignment tools are evaluated based on some proposed tests and obtained results. We will show the tools presented in the last edition in 2016, but first we are reviewing some important tools presented years before.

- AgreementMakerLight, (Faria et al, 2013), it is based on a previous framework called AgreementMaker and takes its advantages focusing on the efficiency and the management of very large ontologies.
- LogMap, presented in (Jiménez-Ruiz & Cuenca, 2011), is based on the use of logic based heuristic techniques.
- In (Djeddi & Khadir, 2013), XMap uses Artificial Neural Network in order to combine several different metrics into a unique one used to obtain the ontology alignments.
- MAMBA[28] developed by University Mannheim in 2015 uses poses ontology alignment as an optimization problem where to use Markov Logic.

[27] http://oaei.ontologymatching.org/

[28] http://web.informatik.uni-mannheim.de/mamba/

Taking a look into OAEI last edition in 2016, we can find new tools:

- (Khiat, 2016), exposes the results of CroLOM which uses a translator like Yandex[29] and applies NLP techniques and similarity computation to words and its synonyms.
- DisMatch, is presented in (Rybinski et al, 2016), in this tool the main input is the application of the Similarity Flooding algorithm.
- PhenoMM, PhenoMF and PhenoMP are different version of PhenomeNET, (Rodríguez-García et al, 2016), an ontology integrating phenotype ontologies or a database of gene-phenotypes associations.
- Linear Program for Holistic Ontology Program (LPHOM), (Megdiche et al, 2016), ontology alignment is modelled by adding linear constraints to the maximum-weighted graph.
- In (Zhao & Zhang, 2016), is presented FCA-Map based on the mathematical model Formal Concept Analysis (FCA).
- Finally, ALIN is an interactive ontology alignment tool using WordNet[30] as external source, (da Silva et al, 2016).

2.7 Data Retrieval.

As we have said before there is an interest in knowing the structure of the Web of Linked Data as this can be used to improve data retrieval strategies. Data retrieval is the process of identifying and extracting data from a resource containing structured data, for example a database. A mathematical definition is presented in (Thanh, 2011):

Definition 14. The data retrieval model is a tuple $\{R,Q,M(Q,R)\}$ where

1. R, the resource model, comprises of structured data.
2. Q, the query model, is a set of structure constraints defined on the results using a structured query language.
3. $M(Q,R)$ is the framework for matching the structure constraints Q against the structured data R. In particular the matching function $M: Q \times R \rightarrow \{0,1\}$ outputs whether resource in R is a result to a query in Q or not.

[29] https://translate.yandex.com/

[30] https://wordnet.princeton.edu/

Data retrieval should not be confused with information retrieval, which is the same process for unstructured data. The problem sometimes arises because in the literature we can find different classifications and definitions for the different types of retrieval in computer science. For example, information retrieval could be the general discipline that encompasses data and document retrieval, being document retrieval understood as the process of extracting information from documents. But if documents are unstructured data, document retrieval and information retrieval the same can be considered the same. This is because the research community working in information retrieval has mainly work with documents, so most of the time when talking about information retrieval, they are referring to document retrieval. So, summarizing, data retrieval is related to work with structured data and information retrieval to work with unstructured one. There are more differences that can be found in Table 2.2, included in (van Rijsbergen, 1979). Finally related with our research we can talk about semantic data retrieval, as almost every resource in Semantic Web stores structured data.

	Data retrieval	**Information retrieval**
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Classification	Monothetic	Polythetic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive

Table 2.2: Information retrieval vs Data retrieval.

Now we are explaining the differences remarked in the table. When we talk about matching, in data retrieval the user is interested in knowing if a particular item is stored or not. In information retrieval, a set of similar results will be retrieve and then those with the best match will be chosen.

Inference in data retrieval is deductive; if *"a"* is related with *"b"* and *"b"* is related with *"c"* then *"a"* is related with *"c"*. In the case of information retrieval, we have to talk about inductive inference, relations are specified with a degree of certainty or uncertainty and hence our confidence in the inference is variable.

Related with the previous point is the model used to do the inference. In the case of data retrieval, it is deterministic based on the relations among states and events. In information retrieval, the model is probabilistic and applies Bayes Theorem.

In terms of classification, we have to distinguish between monothetic and polythetic strategies. Monothetic is used in data retrieval; it is based on the idea that an item is part of a class depending on the value of a single variable. In polythetic classification, used in information retrieval, this is made based on various variables.

The query language will be artificial, using a syntax and a vocabulary, for data retrieval and natural, the one used by humans, for information retrieval.

The query specification is related with the information we want to retrieve. As in data retrieval we are interested in extracting a particular piece of data, the query must be complete. In information retrieval, the query could be incomplete as the aim is to obtain relevant documents. As seen here, the items wanted are an exact match in data retrieval and relevant items in information retrieval.

Finally, exists the error retrieval, which is directly related with the information above. In data retrieval, we need to obtain an exact match and we are working with a syntax and a vocabulary, due to that a simple error gives us different results or even an inconsistency. In information retrieval, an error can give us different results but also valid.

As said before data retrieval can be classified as semantic data retrieval or Semantic Web Retrieval (SWR), which includes ontology search, linked data search and others. In the following Figure, described in (Butt et al, 2015), we can see a complete process of SWR.

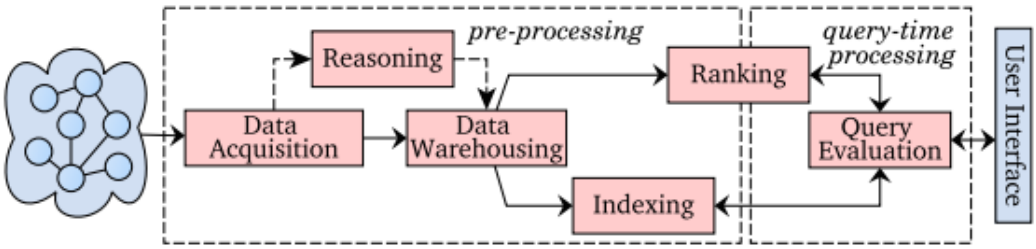

Figure 2-8: Semantic Web retrieval process. (Butt et al, 2015),

Regarding (Butt et al, 2015), the process of data retrieval is complex and has been divided in several steps. The following figure obtained from the previous paper explains the process. In the figure, the boxes are the steps of the process and the arrows how the data flows. The first step “Data Acquisition”, one of the most important, consists of using structured Semantic Web data crawlers for crawling data, (Van de Maele et al, 2008) or (Isele et al, 2010). The aim is to obtain linked data as quick as possible and in an efficient way. Then “Data Warehousing” is used to define which data the user is interested in, the automatization of its extraction, transformation and load. Between the first and the second step, we could find the “Reasoning” process. This step is necessary because sometimes data from the crawled data is inferred using reasoners, for example (Haarslev & Möller, 2003) or (Glimm et al, 2014). Once the data has been stored the process continues by giving a result to the user, as the amounts of data are very large and the infeasible times of response. To solve this problem, data is stored using an URI called key and it is also decided where to store the information in the disk. These kinds of techniques are called “Indexing” or “Ranking”, the difference is that the second one tries to give the most appropriate for the user query. After that, the data will be available to be retrieved. To access the data, applications provide a user interface where users write their queries. These queries go through a process of validation after accessing the data that the user wants to retrieve.

We also are taking into account (Butt et al, 2015) regarding a classification of SWR techniques. We find five main categories, which are subdivided in several subcategories. If we add all the subcategories, there are in total 16 dimensions covering different techniques. A short definition of their dimensions is given below so it could be better understood how they have been grouped:

- The first categories are related with retrieval design decisions: there are techniques depending on the data the user is interested in, how the user initiate the retrieval process or the type of results that can be obtained.
    - Scope: if the user is interested in the classes and properties, we talk about Ontologies. If the user is more interested in the entities, their relationships and subgraphs, these are Linked Data. Finally, there are techniques mostly applied to graph-based data called Graph structure data.
    - Query model: keyword search where a user makes a query composed of one or more keywords and the given results include one or more of these keywords. Structured query search, based on a syntax the user performs a more complex query, which will give more accurate results. Faceted browsing lets users to filter results. Finally, hyperlink navigation is used to navigate within data by clicking the different hyperlinks.
    - Results type: relation centric, find relationships between entities. Entity centric, by searching in several documents it compiles different information about an entity and presents it as a profile. Document centric, it provides a set of URIs or labels of matched documents or parts of them.
- The second category is called Storage & Search, related with the process of retrieving and storing the data.
    - Data Acquisition: this classification depends on how the data is collected. Mainly, it exists manual collection and crawlers. The first one is made by a user himself taking into account the data he needs. Crawlers are applications design to gather data automatically in an efficient and fast way. The crawlers can be classified into: HTML agnostic crawlers which do not crawl HTML documents, HTML aware crawlers working with RDF and HTML documents, and focused crawlers, which limit the use of HTML in order to focus in RDF data.
    - Data Storage: as its name says is based on how information is stored. Relational Databases are used to store triples or quads as in the traditional databases. NoSQL Databases are also used in order to improve data processing and storage. Finally, Native Storage uses their own architecture to store information.
    - Indexing: covering the techniques used to index Linked Data, four categories can be found. Full text index, implemented as an inverted composed of a

lexicon. Structural index, normally used for RDF storage with structures of triples or quads. Graph index, used over graph or RDF data. Multi-level indexing, creates a multiple type of index SWR techniques.

- Query Match: depending on how accurate the queries must be. Exact Match only gives results when the results satisfy all the conditions in the query. Partial Match, is based on the fact that the user is interested in a set of data that accomplishes some of the conditions of the query.

- The third category is Ranking, as some techniques establish a ranking where the result in first position seems to be the best for the user. It also has several subcategories depending on the technique.
  - Ranking Scope is used to denote if a technique depends on a query or not. There exists Global and Focus approaches. In Global the ranking is applied to the whole dataset where the query is used. Focus is when the ranking is applied to the results obtained from the query.
  - Ranking Factor is based on how the ranks are calculated and it also covers different subcategories. Popularity, like PageRank or Tf-idf where the ranking is based on how an item is important respect to the rest of a set. Authority, based on a factor of trustworthiness. Informativeness is based on the idea of how much an item is described to be considered as unique. Relatedness, where the positions in the ranking depend on how similar two items are. Coverage is related with queries, establishing how much they are covered by a resource. Learning model is based on machine learning, choosing some features a ranking is produced and then will be learnt in order to produce other rankings. Centrality, applying the idea to ontologies or graphs where a node with a high connectivity is more important. Finally, based on the users' opinions, it exists the User feedback. It also must be taken into account that there are techniques not using ranking, catalogued as No Ranking.
  - Ranking Domain, here are classified the ranking techniques depending on their domain. We have techniques from the Semantic Web domain. Graph database as the RDF model is represented as a graph. Document retrieval from where most of the techniques have been adopted. Finally, Machine learning, as this field has also been applied for rankings.

- Next category called Evaluation as the different techniques need to be evaluate by three factors: efficiency, effectiveness and scalability.
    - Efficiency is evaluated with: query execution time, time execution time and index update time.
    - Effectiveness uses several metrics: Recall, number of useful documents that are retrieved; Precision, fraction of retrieved data that is relevant; F-Measures based on precision vs recall; Mean Average Precision (MAP), the average precision of a query over all the run queries and Normalized Discounted Cumulative Gain (NDCG), is a standard evaluation measure for ranking tasks with a non-binary relevance judgement.
    - Scalability depending on the size and complexity of datasets or queries: data size, data complexity, query size and query complexity.
- The last category depends on the Practical Aspects and it is divided in three categories.
    - Implementation covers de developing language used like: Java, Python, C# or C.
    - Datasets depends on if the dataset contains real-world information or if they have been developed synthetically. Then we have Real and Synthetic categories respectively.
    - User Interface, if the user interacts through a graphic user interface with the application it is a GUI and if it uses Web services is an API.

To summarize the different categories and subcategories and obtain an idea of how the categories could be classified and grouped, we have Figure 2-9 below.

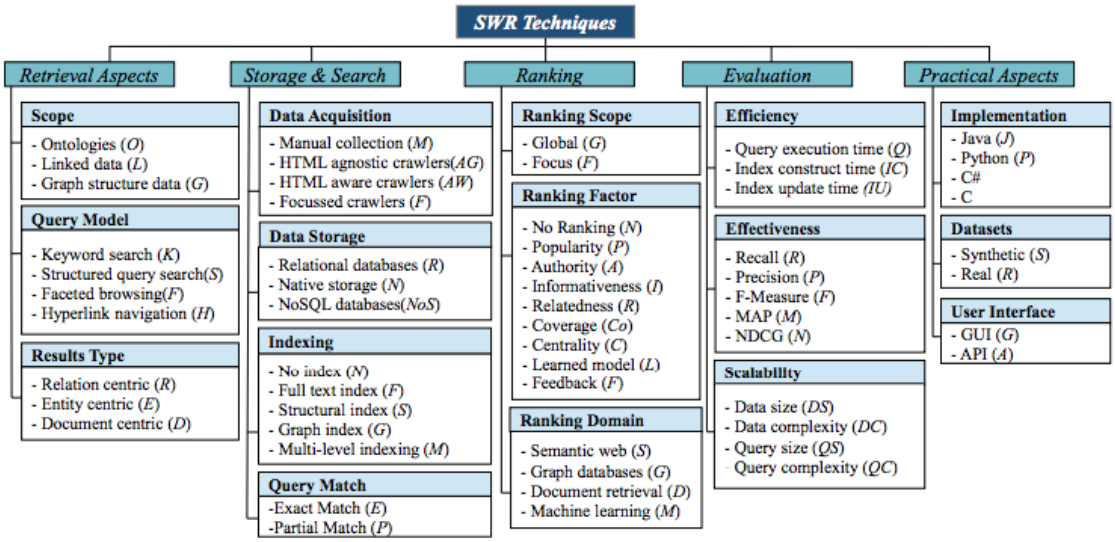

Figure 2-9: Semantic Web retrieval techniques by dimensions. (Butt et al, 2015).

In order to better understand the previous classification, we present some papers approaching the categories listed above. We will start from the **Retrieval Aspect** techniques depending on their *Scope*. In (Cantador & Castells, 2007), a new "ontology" retrieval technique is presented for collaborative reuse and evaluation of ontologies. Using keyword-based search instead of retrieving documents the process will retrieve ontologies. The improvement of a model based on "linked data" retrieval called Semantic Linked Data Retrieval Mode (SLDRM) is made in (Tran & Nguyen, 2016) by defining a mapping logic structure. The mappings are established between DBpedia, using some IDs to Wikipedia articles, and YAGO2[31] by making links to retrieve entities from YagoFacts, YagoLiteralFacts and YagoTypes. Finally, "graph" retrieval is used in (Lux & Granitzer, 2006), where the authors take advantages of the suffix tree model. In the paper, the graphs are comprised as trees so the storage data size is reduced making possible faster retrievals.

In the *Query Model* group, we can find (He et al, 2007), a "keyword-based" SWR technique applied on data graphs. In (Yuan & Mitra, 2013) is presented Lindex, a graph index for database "graphs indexing" subgraphs contained in them. BioPortal is a biomedical repository of ontologies with a Web interface that allows to "browse", search and visualize ontologies, (Noy et al, 2009). Finally, we have the "hyperlink-based" techniques, for example

[31] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

in (Vesse et al, 2010) is developed a method for locating linked data by adapting ideas from hypermedia link integrity.

Next category is *Result Type*. Based on "document centric", (Guo & Heflin, 2007) has built a knowledge base system for OWL documents. In (Dietze, 2016), a discussion based on results is used to improve search and retrieval "entity centric" techniques. The last approaches are "relation centric" techniques, Cheng et al 2014) describes a tool to exploratory association search called Explass.

In the second main category, **Storage & Search**, we can find techniques related with *Data Acquisition*. Here are previous researches using techniques of "manual collection" like (Lizio et al, 2015), where is presented FANTOM5[32] a collection of human and mouse cells that are in part collected manually. In the group of the "HTML agnostic crawlers", we can find a discussion of the implementation and architecture of Semantic Web Search Engine (SWSE) in (Hogan et al, 2011). In (Emandadi et al, 2014), we can find an approach of an "aware HTML crawler" where it is calculated the probability that an HTML link leads to an RDF document in base to prioritize them. The last category is about "focused crawlers", in (Jain & Rawat, 2013) a survey comparing different crawlers is presented.

In the *Data Storage* category, we can find "native storage" like (Furche et al, 2010) that presents G-Storage a lightweight storage manager for graph data. (Cudré-Mauroux et al, 2013) makes a comparison of "NoSQL store" for RDF data. In (Finin et al, 2005), Swoogle[33] a Semantic Web search engine storing the information considered in "relational databases" and categorized as such.

In the *Indexing* category, we can find the "full text index" techniques with papers like (Minack et al, 2008) where an Information Retrieval approach combining structured RDF queries and full-text indexing is described. "Structural index" is represented in papers like (Fox et al, 2014), in which a structural index and a query optimizer that does not use joins operations is proposed. A "graph index" for large volumes of RDF data, is described in (Udrea et al, 2007). "Multi-level indexing" combining trees, hashing and matrices is proposed in (Sankar et al, 2014). Apart from that, we have to consider implementation with "no index".

[32] http://fantom.gsc.riken.jp/5/

[33] http://swoogle.umbc.edu/2006/

The last subcategory in this category is *Query Match.* Here, *"Exact Match"* is used in OWLS-MX a hybrid matchmaker using logic-based reasoning, (Butt et al, 2014), is applied partially matching in an algorithm called DWRank used to rank concepts in ontologies.

The next main category is **Ranking**, whose first grouped techniques are related with *Ranking Scope*. AKTiveRank, (Alani et al, 2006), is a system that ranks ontologies based on a set of metrics; it could be classified as *focused ranking*. Using *global ranking* techniques, we have (Ning et al, 2008) whose approach is Really Simple Syndication[34] (RSS), a framework for searching semantic data resources.

The *Ranking Factor* category has subcategories as "popularity"; whose most famous algorithm is maybe PageRank, applied to Semantic Web in (Lamberti et al, 2009). Talking about "authority", HITS algorithm is used in (Franz et al, 2009). "Informativeness" is used in (Meymandpour & Davis, 2013), where a novel approach derived from information theory is proposed, aiming to measure informativeness in the field of Web of Data. "Relatedness" is a measure used in well-known tools, for example WordNet, (Fellbaum, 1998). "Coverage" ranking, is applied in (Turney & Pantel, 2010) with the Vector Space Model. The technique of "learning to rank" is used in (Butt et al, 2016), where an algorithm to rank concepts in ontologies called DWRANK is presented. Rankings as "centrality" can be found in (Zhang et al, 2007) for ontology summarization. Finally, there are researches whose ranking is based on the used "feedback".

The next category called *Ranking Domain*, it comprises papers with rankings designed for the "Semantic Web" or brought from other fields. Ranking from "graph databases" is found in (Alahmari &Thom, 2015), where a ranking based on the importance of attributes computes the shortest path between nodes. From "document retrieval", we have found BM25F an extension of a previous ranking, (Perez-Agüera et al, 2010). Related with "Machine Learning", in (Arora & Vikas, 2011) some approaches from this field to rank are compared.

The next big category is **Evaluation**. First subcategory is "Efficiency", with works like: (Roatis, 2014) where is measured the *query evaluation time,* (Li et al, 2010) measuring the *index construction time* or *index update time*. The rest of the categories "Effectiveness" and "Scalability" are not going to be reviewed here because they are based in basic characteristics, being difficult to find papers related only with them.

[34] http://www.rss.nom.es/

The final main category is **Practical** aspects whose first subcategory is based on *implementation*, depending on the programming language used. Java is used in (El-Sappagh & Elmogy, 2016), Python in (Schiessl et al, 2017), C# in (Yazhmozhi et al, 2013) or C in (Sun et al, 2013). Regarding the *type of datasets* used in the research, there are: (Schmidt et al, 2011) with “synthetic data” or “real data” from DBpedia like (Fafalios et al, 2016). Finally, there are researches depending on the *user interface*: in (Zhang et al, 2009) using a “GUI” or in (Chen et al, 2008) using “APIs”.

3. STUDIES

The next section describes the studies that have been conducted during the research period, in order to accomplished for the research objectives discussed in the previous section.

3.1 Data Retrieval from the Web of Linked Data

In this section, we are describing step by step the experiment achieved to accomplish with O1: “Connect the Web of Linked Data with an independent data source”.

3.1.1 Motivation

The main question about this research is related with the structure of the Web of Linked Data. As we have said before the Web of Linked Data has to be seen as the Web of Documents but instead of having Websites, we navigate through datasets containing information in different fields. By having an idea about its structure, we can relate it on how the data retrieval strategies are designed.

But the first issue, is to prove that the Web of Linked Data is accessible to crawl data. Normally, this information can be accessed through SPARQL endpoints or by downloading a dump of the dataset. For both, there is a minimum level of expertise in knowing the SPARQL syntax or the structure of the dataset. So, first we need to create a process that user could automatically retrieve some content from the Web of Linked Data.

At this point, we need to find some independent data source that could take advantage of using information of the Web of Linked Data. But we also need a way to fill the gap between both data sources. We know that information in the Web of Linked Data is described by vocabularies. For example, if we have information about Madrid, a vocabulary with the term “City” is used to describe that Madrid is a city. There is a need to find data sources using any of these vocabularies or vocabularies being used in a dataset. Then, establishing which vocabulary terms are being used in both data source, we can take information from the Web of Linked Data and use it to enrich the other.

3.1.2 Introduction

The first step in the experiment is finding an independent resource from the Web of Linked Data that lets us to aggregate new information retrieved from it. It exists a vocabulary/ontology called Schema.org that was created by Bing, Google and Yahoo! and launched on June 2, 2011. Schema.org ontologies are intended for the creation of

microcontents targeted in improving indexing and search systems, (Johnsen, 2012). This vocabulary consists of a set of tags defining terms that could be used in HTML5, so Webmasters could mark-up their Websites with microdata. Microdata helps search engines and other tools used in Websites to better understand the information used in them. By tagging the Websites results given by search engines will be more accurate. The importance of using Schema.org can be reviewed in (Mika & Potter, 2012). Taking that into account, we can use any Website using Schema.org as independent resource.

Now that we have located the independent source of data to be populated with information from the Web of Linked Data, we need a process to retrieve the data to be added. We know that information in the Web of Linked Data is described by vocabularies, so a good solution will be to find how to get to these vocabularies. This can be achieved with LOV, as we know it is a catalogue that comprises all the vocabularies used in the Web of Linked Data. So, if we are able to create links between Schema.org and LOV, will be able to bring the information from Web of Linked Data to Websites. As we can consider Schema.org as an ontology, the process to obtain the links consist of applying ontology mapping techniques.

3.1.3 Materials

As said before, our starting point is the Schema.org vocabulary that can be found in the following formats: microdata, RDFa[35] and JSON-LD[36]. Schema.org has evolved from the 302 classes and 286 properties in the first release to 603 classes and 851 properties in 23rd of March 2017. An evolution of the number of classes and properties through its releases can be seen in the following Figure.

[35] https://rdfa.info/

[36] https://json-ld.org/

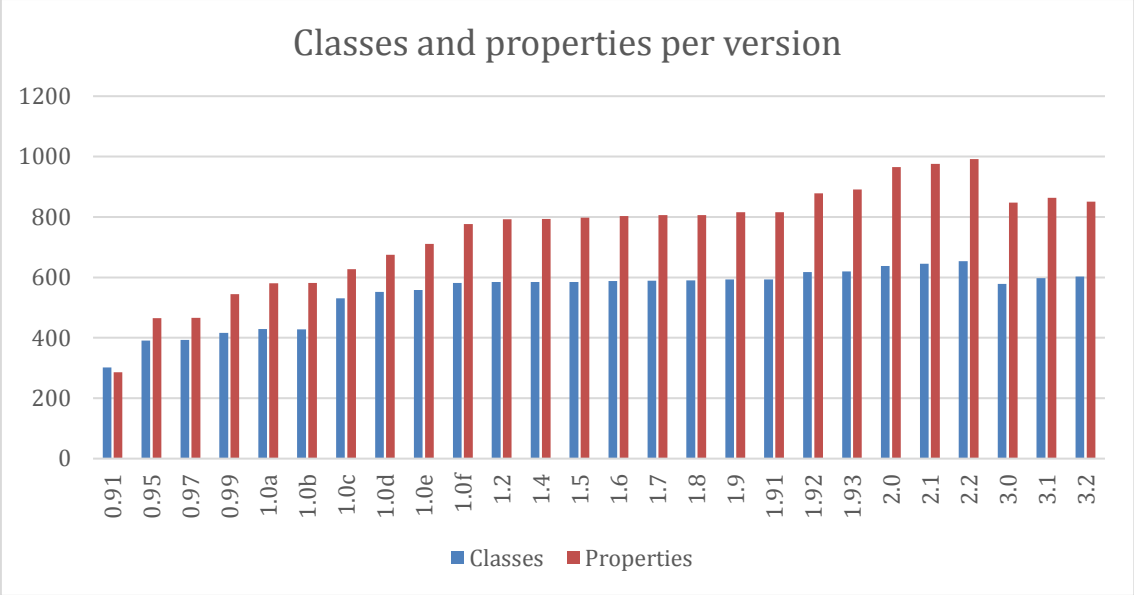

Figure 3-1: Evolution of Schema.org.

It has the form of a hierarchical tree of classes, having each class a set of properties. The broadest class is “Thing” whose properties are: name, description, url and image. Due to its structure items inherit properties from their parents, for example: Book as a narrower class from CreativeWork. There is also a classification for the data types, see Figure 3-2. “Thing” class is divided into 8 main categories which are: “Action”, “CreativeWork”, “Event”, “Intangible”, “Organization”, “Person”, “Place” and “Product”. A distribution of these classes can be seen in Figure 3-3.

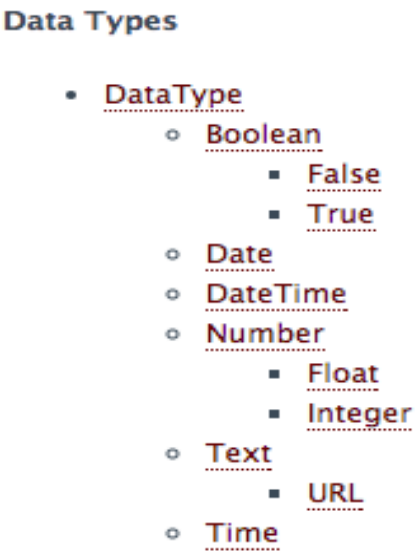

Figure 3-2: Classification of Data Types.[37]

[37] http://schema.org/docs/full.html

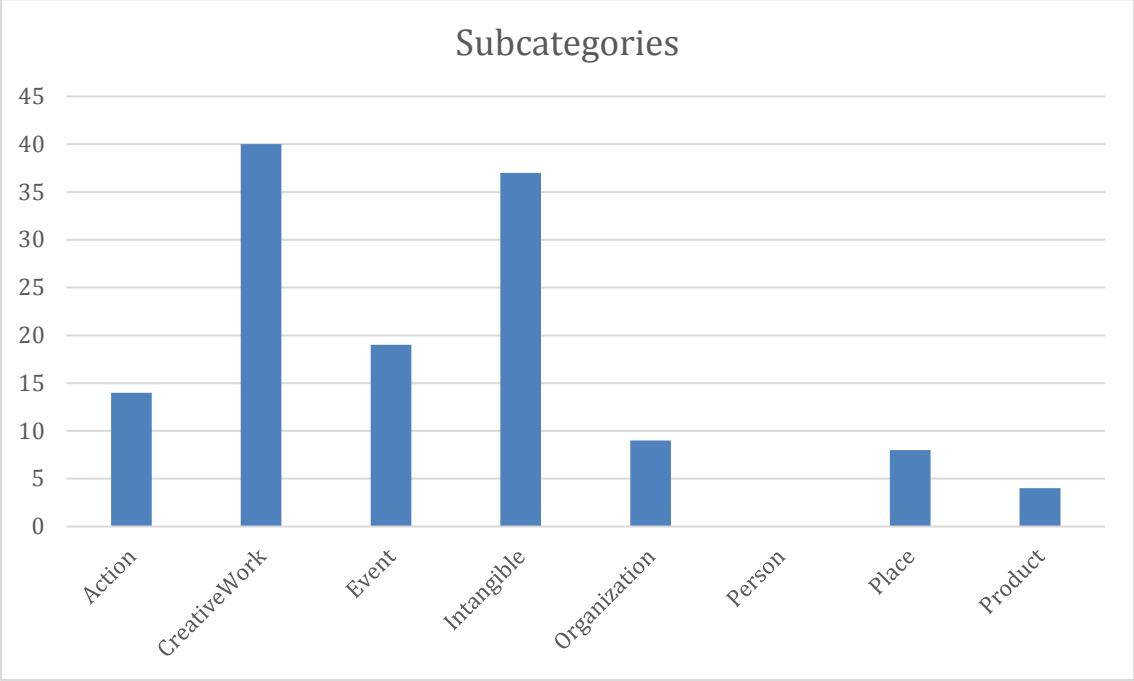

Figure 3-3: Distribution of main categories.

The experiment consisted on retrieving data from the Web of Linked Data by using ontology mapping techniques. For that purpose, we will take advantage from the project LODStats. It is defined as an approach to generate statistics from RDF datasets, (Auer et al, 2012). It gives us a comprehensive picture of the Web of Linked Data. It is comprised by 9,960 datasets with the condition that 6,971 of them are giving problems. From those that work properly, 2838 are accessible with a dump and 151 via SPARQL endpoints. Other overall statistics are: 192,230,648 triples, 3,840 classes, 49,916 properties, 2,593 vocabularies and 1,845 languages. The next three tables show the top 5 used vocabularies, classes, properties and languages.

Vocabulary	Occurrences
http://www.systemone.at/2006/03/wikipedia	35,310,770
http://www.w3.org/1999/02/22-rdf-syntax-ns	25,056,073
http://purl.org/dc/terms/	22,590,888
http://www.aktors.org/ontology/portal	16,672,643
http://www.w3.org/2004/02/skos/core	12,013,166

Table 3.1: Top used vocabularies.

Class	Occurrences
Manifestation	2,492,000
BibliographicResource	2,434,091
Article	2,049,423
Person	1,991,834
Book	1,903,547

Table 3.2: Top used classes.

Property	Occurrences
internalLink	34,449,952
type	22,697,337
label	4,042,604
identifier	3,949,709
subject	2,870,884

Table 3.3: Top used properties.

Language	Occurrences
English	1,598,183
Italian	1,117,198
French	666,932
Deutsch	622,162
Spanish	24,918

Table 3.4: Top used languages

The next step consists of linking both data sources: Schema.org and LODStats by finding mappings with LOV. LOV is an observatory of a catalogue of vocabularies used in linked data. Starting in 2011, it has grown till the amount of 603 vocabularies in July 2017. Its purpose is promoting and facilitating the access to the vocabularies, describing relations between them and how they are connected to the Web of Linked Data. The information in LOV can be retrieve with an SPARQL endpoint or using they LOV API[38]. It also provides a dump in n3 and nq.

To achieve the aims described above, LOV provides the following tools:

- **Ontology Search**: LOV provides the search of information between vocabularies. Users can find vocabularies, terms or agents (people responsible of a vocabulary).
- **Ontology Assessment**: LOV supports ranking classifications. In total eight different methods grouped into two main categories; Tf-idf, BM25, Vector Space Model, Class Match Measure, PageRank, Density Measure, Semantic Similarity Measure and Betweenness Measure.
- **Ontology Mapping**: relations between vocabularies are described by the usage of Vocabulary of a Friend (VOAF) vocabulary[39]. The relations are described by the following properties: reliesOn, usedBy, metadataVoc, extends, specializes, generalizes, hasEquivalencesWith, hasDisjunctionsWith and similar.

[38] http://lov.okfn.org/dataset/lov/api

[39] http://lov.okfn.org/vocommons/voaf/v2.3/

For each vocabulary, it provides a dump file in n3. A set of external tools to visualize or interact with it like: triple-checker[40], VAPOUR[41], Parrot[42], OOPS![43] and WebVOWL[44]. Also, general characteristics like: classes, properties, datatypes, instances, URI, namespace, description, language or creator. Finally, a graph showing how vocabularies are related by showing its incoming and outgoing links. The two following Figures shows this information.

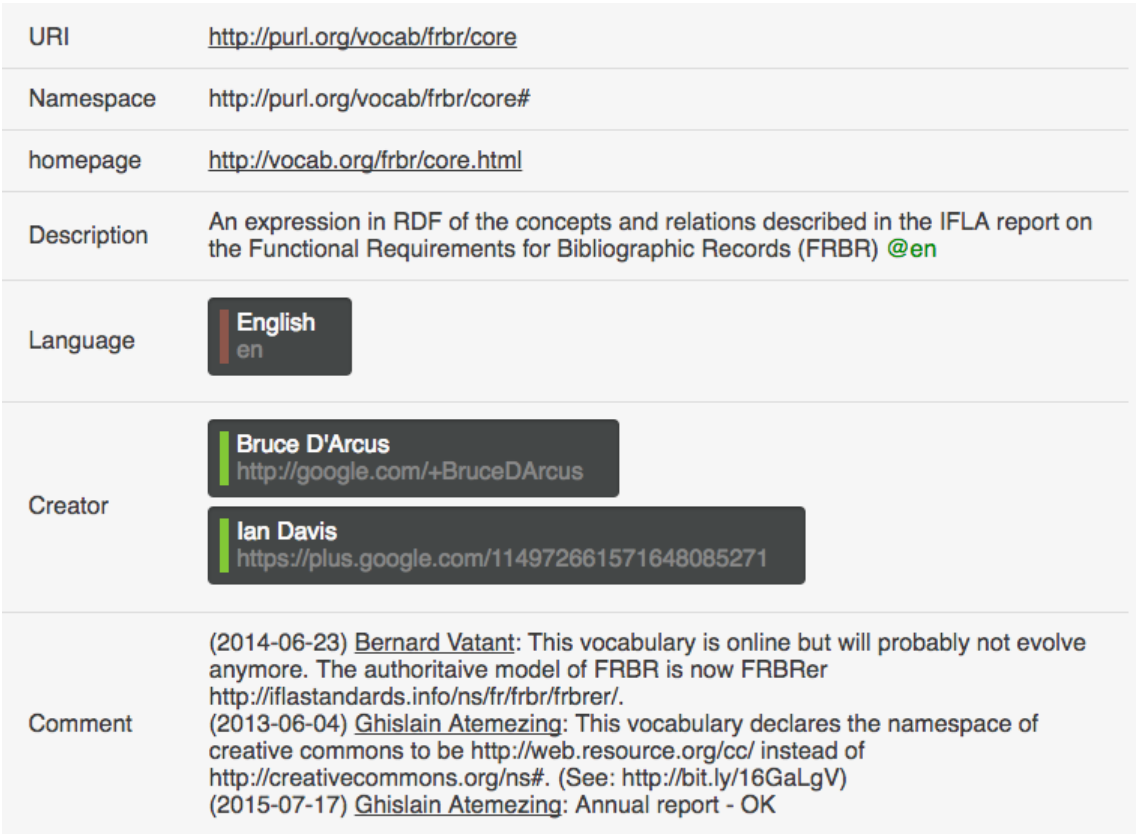

Figure 3-4: General characteristics of a vocabulary in LOV[45].

[40] http://graphite.ecs.soton.ac.uk/checker/

[41] http://vapour.sourceforge.net/

[42] http://ontorule-project.eu/parrot/parrot

[43] http://oops.linkeddata.es/index.jsp

[44] http://visualdataweb.de/webvowl/

[45] http://lov.okfn.org/dataset/lov/vocabs/frbr

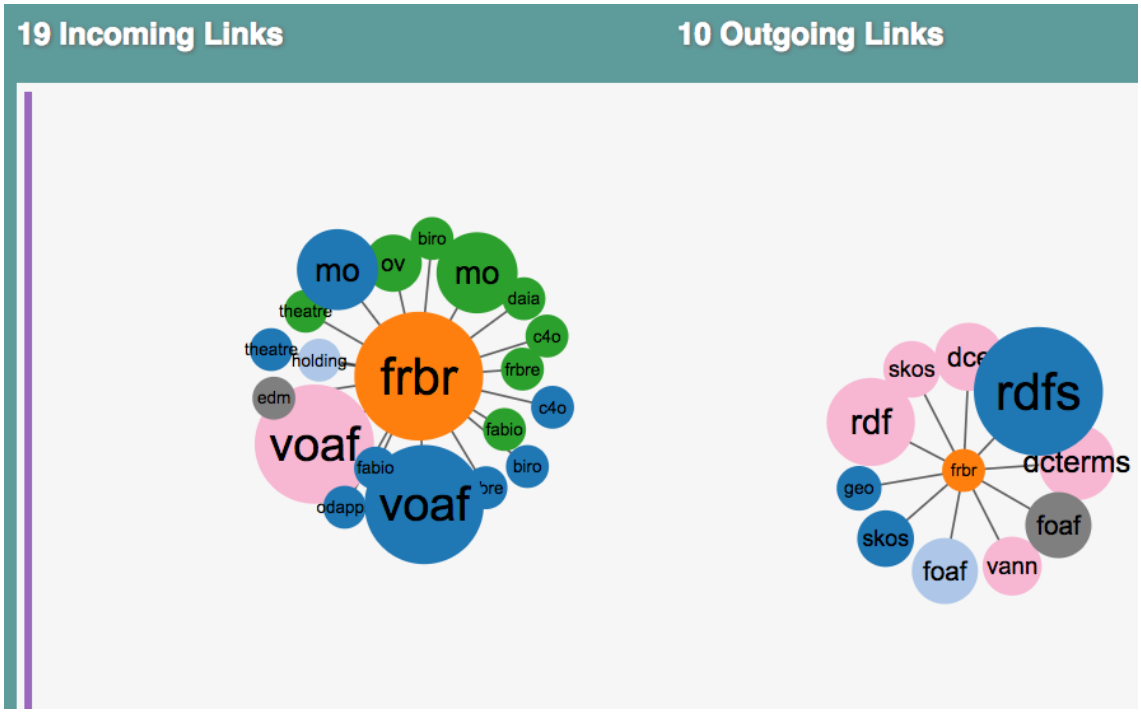

Figure 3-5: Graph showing links between vocabularies[46].

At the moment of the experiment, September of 2013, Schema.org had 429 classes and 589 properties, LOV was formed by 360 vocabularies and LODStats 2289 datasets.

3.1.4 Method

To accomplish the first objective, we will try to stablish a set of mappings between Schema.org and LOV. Then we will measure the impact of this mappings in the Web of Linked Data with the statistics provided by LODStats. To obtain the mappings we will develop a script using ontology mapping techniques. In the Following figure is shown the workflow of the method.

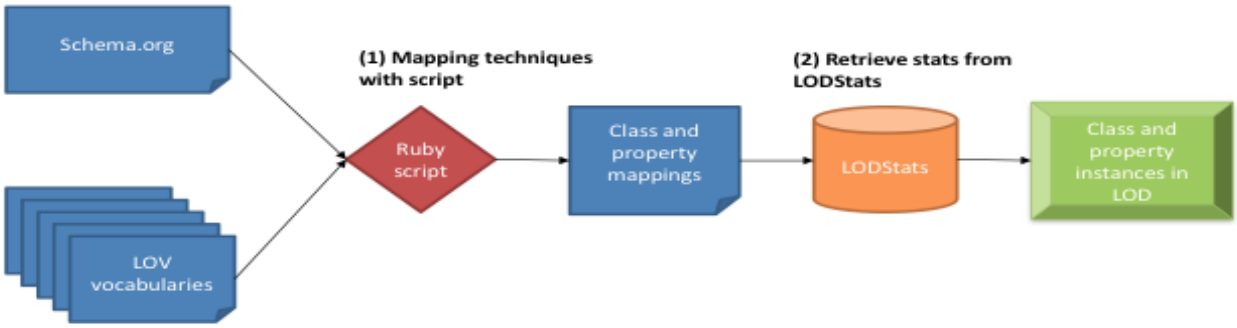

Figure 3-6: Workflow for mappings.

[46] http://lov.okfn.org/dataset/lov/vocabs/frbr

The mappings can be of two types: at syntactic level and at semantic level. A mapping at a syntactic level means that two words are the same if they are spelled in the same way. A mapping at a semantic level is when two words have the same meaning, this is the case of synonyms. Also, the mappings are obtained in two steps, first we are trying to find mappings between classes. Then if a class has a mapping we will try to find mappings for its properties. This means that if there is a mapping for the class "Person" and this class has a property called "familyName", there will be a mapping if this property is also part of the vocabulary a LOV for the class “Person”. The following tables show examples for both.

Class from Schema.org	Class from foaf vocabulary
http://schema.org/Person	http://xmlns.com/foaf/0.1/Person

Table 3.5: Example of class mapping between Schema.org and a LOV vocabulary.

Class from Schema.org	Property from Schema.org	Property from foaf vocabulary
http://schema.org/Person	http://schema.org/familyName	http://xmlns.com/foaf/0.1/familiyName

Table 3.6: Example of property mapping between Schema.org and a LOV vocabulary.

The script has been developed in Python using the packages rdflib[47] and PyDictionary[48]. The first one is a library for managing RDF that can be in different formats as n3, NTriples or Turtle. The second one lets the user get translations, meanings, synonyms and antonyms from words. First, we will compare terms string by string, for the syntactic level mappings. Then, we will make the comparison with synonyms for the semantic level. With these mappings, LODStats is used to obtain some statistics. This will give us the impact of the mapping term in LODStats.

Finally, the mappings obtained with our method have been compared with some ontology mapping tools. These tools have the aim of obtaining the classes and properties that two ontologies have in common. In this case, we are comparing our results with LogMap, (Jimenez-Ruiz & Cuenca, 2011), which gives results of classes, properties and instances. Also, with Alignment API, (David et al, 2011), an API written in Java to align ontologies.

[47] https://pypi.python.org/pypi/rdflib

[48] https://pypi.python.org/pypi/PyDictionary/

3.1.5 Discussion and results

3.1.5.1 Mappings of classes and properties

Previously, we have distinguished mappings between two different kind of terms. First, mappings are made between classes and then between properties. The information obtained after this, which are the mappings between Schema.org and LOV, will be related with the statistics provided by LODStats. What follows is numeric data from (Nogales et al, 2016) from the top mappings achieved.

As said before, first are obtained the mappings between classes of Schema.org and LOV vocabularies using semantic mappings. In total 135 classes were mapped, which were 25,18% of the classes in Schema.org at the moment of the experiment. Comparing with (Nogales et al, 2013), which was a first approach of this experiment, 16 more classes were mapped. Counting the total number of instances obtained, there are 585. This means 298 more comparing with the foundation paper. To measure the quality of the results, these are compared with the results obtained with LogMap and Alignment API. In the following table are shown the top 5 classes according to the number of instances. The first column corresponds to the results from our script, the second to LogMap and the last one to Alignment API.

Class Name	Script	LogMap	Alignment API
Book	23	17	21
Place	22	19	17
Event	20	13	17
School	20	16	15
Comment	19	17	14

Table 3.7: Comparison of class mappings between our script and two alignment tools.

We have also obtained a histogram in Figure 3-7, showing the concentration of the mapped classes. This tells us if there are a few classes with most of the occurrences or if there are a lot of classes with only a few occurrences. This classification could be used to advise a Webmaster which terms are better to tag their webpages, which will be the ones with more occurrences as they have more impact in the Web of Linked Data. Also, a Schema.org class with more instances in the mappings could take benefits of more LOV's vocabularies.

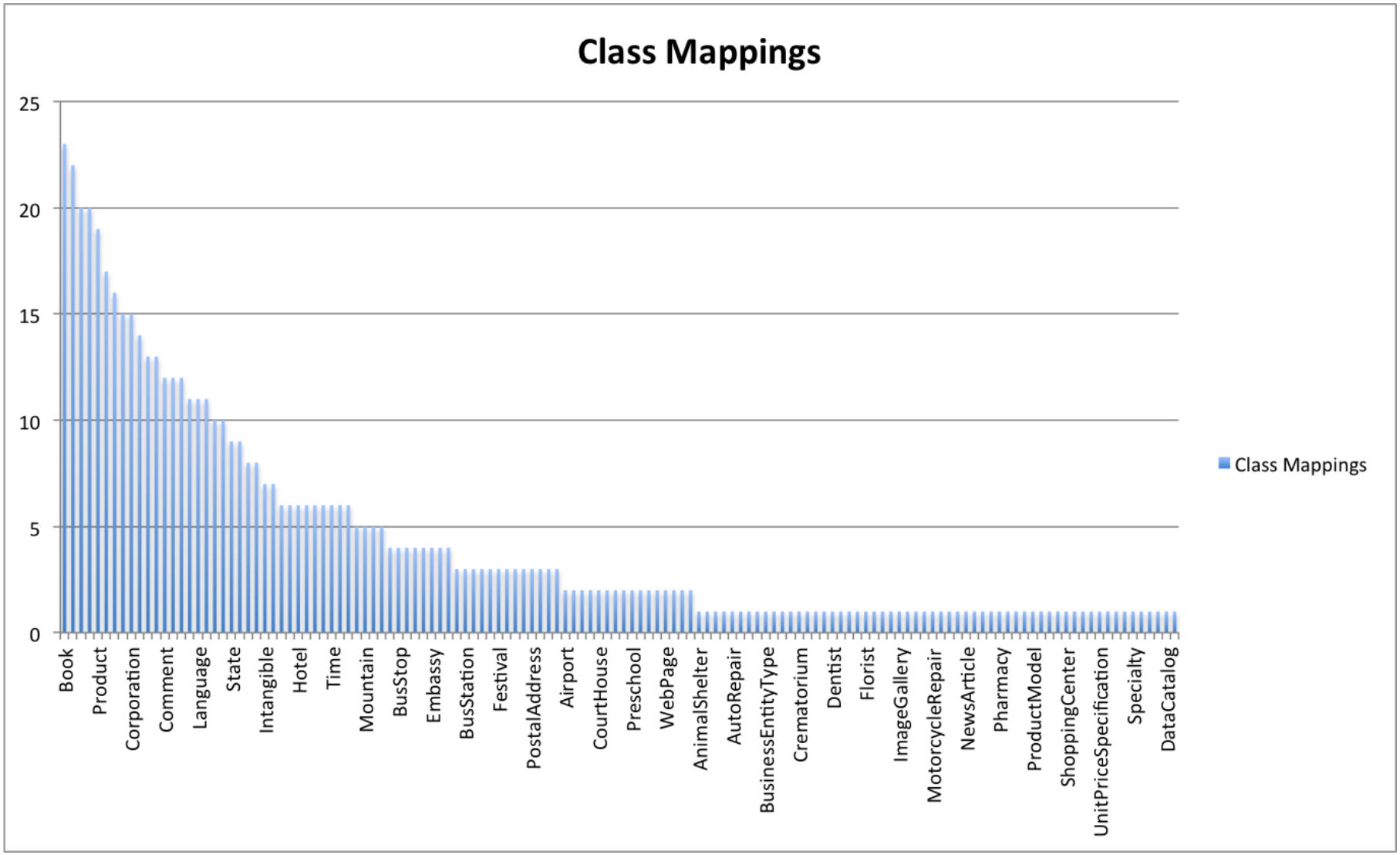

Figure 3-7: Histogram of more classes mapped between Schema.org and LOV.

At this point we have shown the results from the class mappings, now we are doing the same for properties. In this case 16 new properties have been mapped comparing with (Nogales et al, 2013). That means that the addition of the semantic level mapping has improved the results. In total 13,55% of Schema.org properties have been mapped. Counting in total the instances of each property, we have obtained 913. In these mappings 101 different vocabularies have been used. As with the previous results, we show a table comparing our results with LogMap and API Alignment. Also, a histogram to measure the concentration of the mappings can be seen in Figure 3-8.

Class Name	Property Name	Script	LogMap	Alignment API
Table	note	12	11	19
Event	description	9	9	17
AnimalShelter	agent	8	3	14
Winery	agent	8	4	3
Embassy	agent	8	7	11

Table 3.8: Comparison of property mappings between our script and two alignment tools.

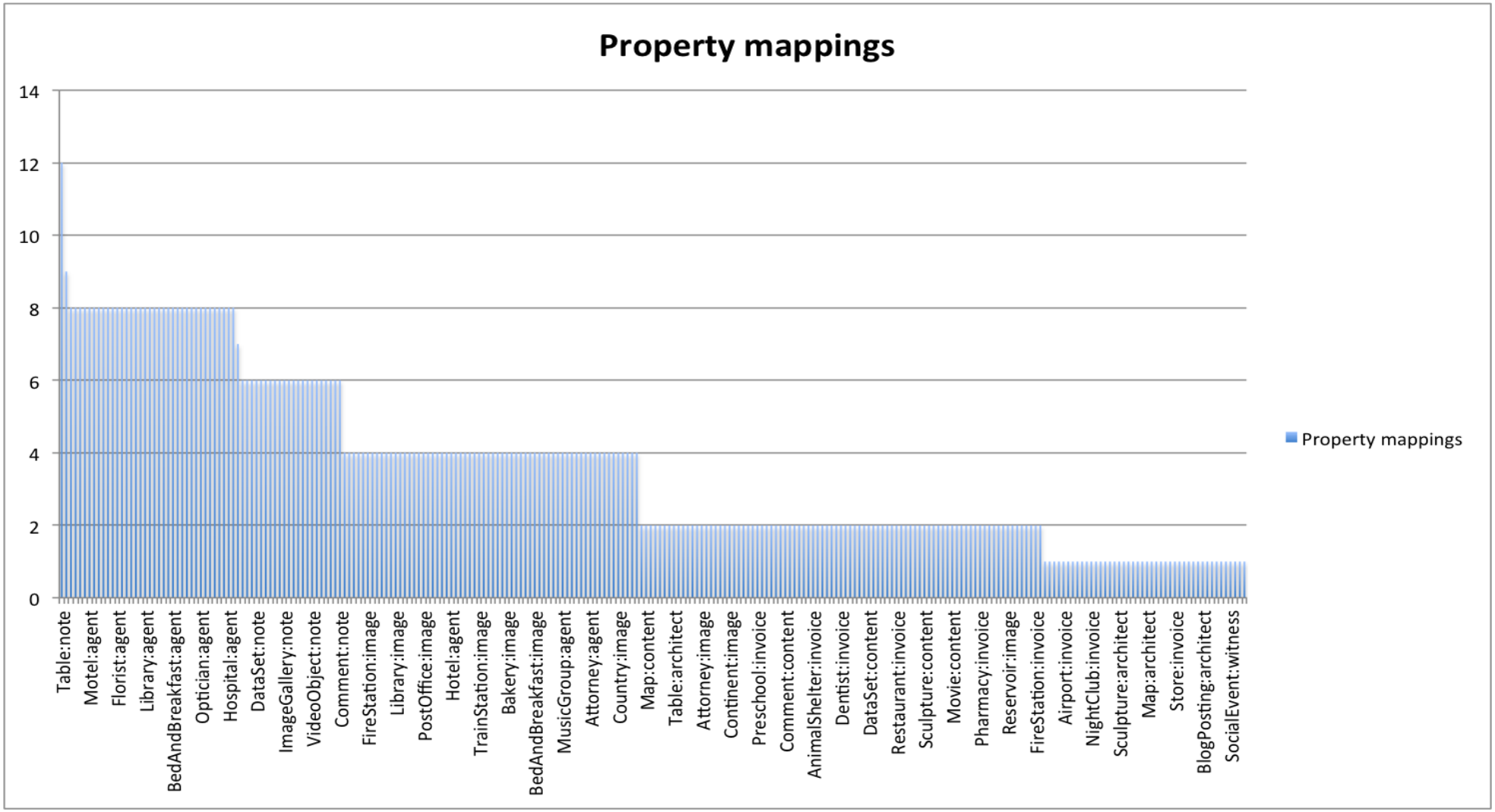

Figure 3-8: Histogram of more properties mapped between Schema.org and LOV.

3.1.5.2 Vocabularies' statistics

Once we have obtained the mappings between classes and properties, it is interesting to give some results about the impact that these mappings have had in the LOV vocabularies. When comparing our results with LogMap, we have found that they sometimes differ. But most of the time, the number of mappings is the same for each vocabulary. Sometimes our mappings have more occurrences and sometimes not. We have also found occasions where LogMap couldn't manage the file, producing an error. In the following Table, we have grouped the different cases giving the percentages for each. In 13 vocabularies, our script obtained better results. LogMap was better in 16 cases. Most of the time the number of mappings were the same, 323 of the vocabularies. Finally, there were 8 cases that couldn't be taken into account as they couldn't be managed by LogMap.

Case	Vocabularies with mappings	Percentage
Script better	13	3.61 %
LogMap better	4	1.11 %
Equal results	335	93.05 %
File error	8	2.22 %
Total	360	100%

Table 3.9: Global comparison between our script and LogMap, classified by cases.

The same comparison but with Alignment API has been made and can be seen in Table 3.10. After running Alignment API with all the vocabularies, we have realised that is less stable than LogMap. In the case of the tool giving an error because it couldn't work with the file, it has occurred 193 times. The error has been a null pointer exception or a problem trying to load some ontologies used in the vocabulary. Taking into account the other use cases: our scripts gave better results in 18 cases, Alignment API in 7 and in 142 cases the results were the same.

Case	Vocabularies with mappings	Percentage
Script better	18	5 %
Alignment API better	7	1.94 %
Equal results	142	39.44 %
File error	193	53.61 %
Total	360	100%

Table 3.10: Global comparison between our script and Alignment API, classified by cases.

There is also an interest in knowing which of the vocabularies from LOV have more occurrences regarding classes and properties. In Table 3.11, this information about the classes of the vocabularies can be seen. In total, about the third part of the vocabularies have a mapping between classes. We have also obtained a histogram to see the concentration that can be seen in the following Figure.

Vocabulary	LOV acronym	Mapping occurrences
Accommodation Ontology	acco	66
LinkedGeoData	lgdo	47
PROTON Extent module	pext	25
Audio Features Ontology	af	18
AKT Reference Ontology	akt	14

Table 3.11: LOV vocabularies with more classes mapped between Schema.org and LOV.

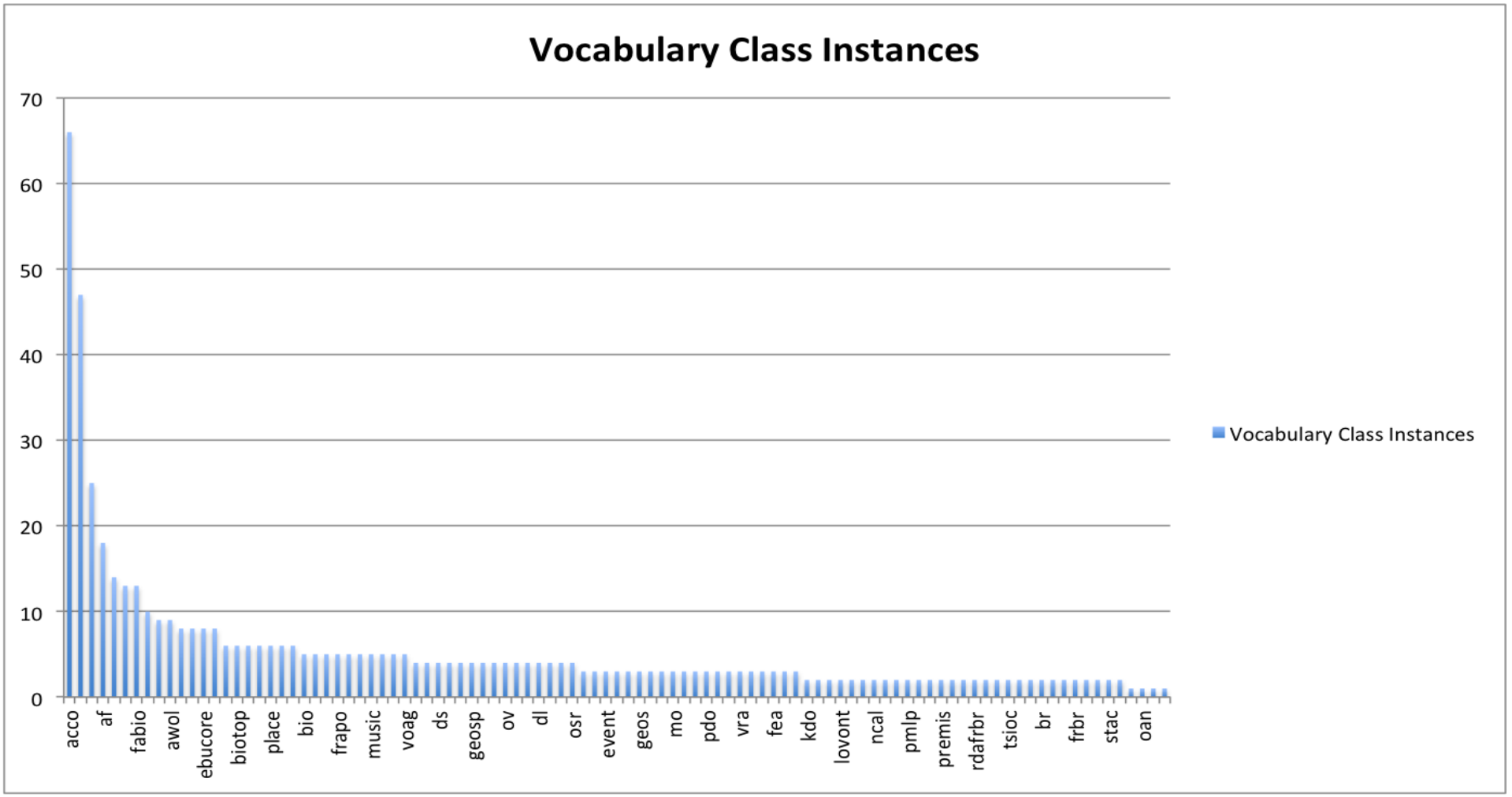

Figure 3-9: Histogram of vocabularies with more classes mapped.

Following, we are talking about the vocabularies in Table 3-12. First one is Accommodation Ontology[49], which is a vocabulary for the description of hotels, vacation homes, camping sites and other accommodation offered for e-commerce. The second ontology is LinkedGeoData[50], a dataset about spatial dimension whose information is collected from OpenStreetMap[51]. The rest of the vocabularies are: the PROTON Extent module[52] which is an Upper-level ontology with extensions to handle Linked Open Data. The Audio Features Ontology[53] is a vocabulary that expresses some common concepts to represent some features of audio signals. The AKT Reference Ontology[54] which describes people, projects, publications or geographical data.

The same Table and Figure but related with the vocabularies' properties can be seen below. The vocabularies found in the Table are: Open Graph Protocol Vocabulary[55] that enables any web page to become a rich object in a social graph. Open.vocab[56], which is a community-maintained vocabulary intended for use on the Semantic Web. BIO[57], is a vocabulary for describing biographical information about people, both living and dead. The Payments Ontology[58], a vocabulary for representing payments, such as government expenditures, using the data cube representation. Finally, the Basic Access Control ontology[59], which defines the element of Authorization and its essential properties, and also some classes of access such as read and write. In total, only 8.05 % of the vocabularies obtained a mapping using the properties.

[49] http://ontologies.sti-innsbruck.at/acco/ns.html

[50] http://linkedgeodata.org/About

[51] https://www.openstreetmap.org/

[52] http://www.bartoc.org/en/node/18173

[53] http://isophonics.net/content/audio-features-ontology

[54] http://projects.kmi.open.ac.uk/akt/ref-onto/

[55] http://ogp.me/

[56] http://vocab.org/open/

[57] http://vocab.org/bio/

[58] https://data.gov.uk/resources/payments

[59] https://bartoc.org/en/node/17815

Vocabulary	LOV acronym	Mapping occurrences
Open Graph Protocol	og	122
OpenVocab	ov	122
BIO	bio	93
Payments ontology	pay	89
Basic Access Control ontology	act	88

Table 3.12: LOV vocabularies with more properties mapped between Schema.org and LOV.

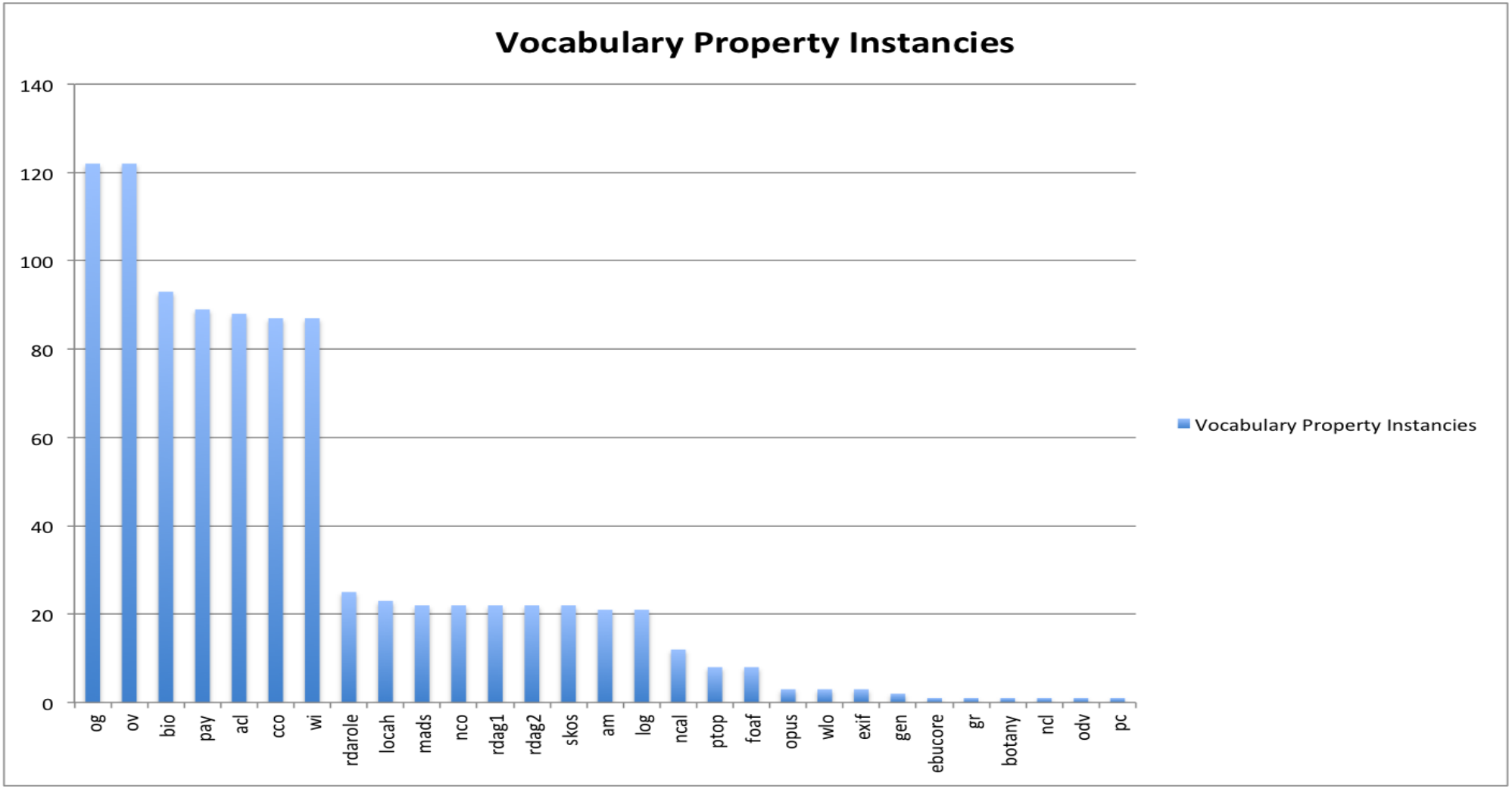

Figure 3-10: Histogram of vocabularies with more properties mapped.

3.1.5.3 Impact in the Web of Linked Data

The second part of our workflow combines the obtained mappings with the statistics provided by LODStats. This will give us an idea about the relevance of the classes and properties in the Web of Linked Data. At this point, we have obtained the number of instances given by LODStats for each class and property that has been mapped. In the following Table, we present the statistics for the classes.

Class Name	Occurrences
Person	3,217,769
Organization	237,655
Event	8,235
City	4,589
Dataset	612

Table 3.13: Schema.org classes from the mappings with more occurrences in LOD.

In table 3.14 we have the same information for the properties mapped between Schema.org and LOV. In this case we have searched in LODStats using only the name of the property, but not considering the previous matches of the classes, as this information is not provided by LODStats.

Property Name	Occurrences
Name	16,656,930
Description	8,784,687
Height	4,718,986
Width	4,718,984
Gender	2,848,501

Table 3.14: Schema.org classes from the mappings with more occurrences in LOD.

3.1.6 Limitations

During the experiment, we have realised that the experiments have some limitations. First of all, LOV has some problems. In the Website where the vocabularies are available for downloading, some of them are not for two reasons. Sometimes, the URL is invalid or there is a problem of content negotiation. Or sometimes, the file that contains the information has never been fetched. Due this, it has not been possible to work with all the vocabularies in the catalogue.

The syntactic mappings are not accurate as good as they could be. When comparing the results of LogMap, we found some special cases. LogMap discriminates symbols like "-". So, in examples like "GovernmentOrganization" and "Government-Organization" they are considered equal. It also considers similar two words that are written different depending on being British English or American English, like "Organization" and "Organisation". Another case is that of synonyms like "School" and "College". Finally, LogMap considers as equal words that contained or are contained by another, like "RecyclingCenter" or "Center". When we are comparing the mappings between our script and LogMap, all these special cases can be taken into account except the last one which we consider an error.

At the point of the mappings, we have a problem with the semantic level making it semiautomatic. When we obtain a mapping like this, we have to check if they can be used in a sentence having the same meaning. This is called disambiguation and has been made manually. To have better results this should be done by vocabulary curators.

There are other limitations in LODStats. The Website highlights that 6971 of the datasets have errors with the dumps or with the SPARQL endpoints. Therefore, we cannot assure that all the information provided is accurate.

3.1.7 Conclusions and outlook

A script has been developed in order to obtain mappings between Schema.org and LOV. We understand LOV as a catalogue of the vocabularies used in the Web of Linked Data, being this a bridge with Schema.org. By using the instances of the mappings in the Web of Linked Data, we can measure the impact of Schema.org in it.

The mappings have been done for to kind of two terms: first to obtain mappings between classes of Schema.org and LOV, and then the same between properties, taking into account the previous ones. The mappings are also made in two different levels: syntactic level that means two are the same if they can be written in the same way and semantic that is the case

that two terms have the same meaning. Ones we have obtained the mappings, we are using LODStats to obtain instances of them in the Web of Linked Data.

By finding the mappings, some statistics have been obtained. The set of mappings between classes is bigger than the mappings between properties. Instead, the instances of the properties in the Web of Linked Data is bigger. Talking about numbers, 135 different classes have a mapping. If we talk about properties, 585 have a mapping.

To measure how accurate our approach is, the results have been compared with two ontology matching tools: LogMap and API Alignment. In the case of LogMap, in 89.72% of the vocabularies we have obtained the same results. API Alignment has given us a lot of errors, so only 53.61% of the vocabularies could be compared with the results of our script.

In future works, these mappings could be used to create new data retrieval strategies. For example, SPARQL-federated queries. Since a federated query provides information from various data sources, we can use mappings to combine the information provided by them.

3.2 Aggregation with data from the Web of Linked Data

In this section, we are describing step by step the experiment achieved to accomplish with O1.1 “Present a use case in which information from the Web of Linked Data is aggregated to an independent resource.”.

3.2.1 Motivation

In the previous experiment, we have been able to connect Schema.org to the Web of Linked Data in order to access the stored data. This is a first approach, so user could take benefit to retrieve information from the Web of Linked Data. The question now is: An important issue will be to design an automatic process so non-expert users could retrieve data from the Web of Linked Data But once this data has been retrieved, another question arises. This question is related about the opportunity of using this data to enrich other data sources.

3.2.2 Introduction

Since the beginning of the digital age in 1990s, there was a need to store the information. At that moment, started the development of tools like the digital libraries, having the aim of storing the biggest amount of information in less space. As people started working on more efficient digital libraries, it also arose the problem of how to recover the data in a more efficient way and wasting the less amount of time. The solution to this problem, is what we mentioned before as information retrieval strategies. The problem of having a good access to the

information means being able to do accurate searches and obtaining the information as faster as it could be. When the Web of documents became a daily used tool, this became also a regular problem for the users. So, when the Web of Linked Data arose it became also a problem.

As we have said before, the field that studies this problem in the Web of Linked Data is called data retrieval strategies. We know that the Web of Linked Data was formed when organizations and companies started to open their data to users. They realised that by setting free the data, users could work with them obtaining new results. Then by adding new datasets that share terms from other datasets, a graph network of datasets was formed. One of the usage that could be interesting, will be retrieving the information stored in a dataset and aggregating it to an independent data source.

For example, this information could be aggregate to Websites. There is a need to know if is possible to connect a Website with the Web of Linked Data. In the previous experiment, we have been able to connect the Web of Linked Data with Schema.org. We have obtained some mappings between Schema.org and the vocabularies used in the Web of Linked Data. So, there is a possibility in bringing information from its datasets to Websites tagged with Schema.org.

Our data retrieval strategy will take advantage by using the mappings obtained before. First, we need a Website using any of the terms found in the mappings. Then, we need a dataset using also a vocabulary with these terms. The most effective way to obtain that, is working with DBpedia as it is the biggest dataset we know. Then, by using the mapping, we can retrieve other information associated to the mapped term and aggregate it to the Website. We will see a formal example of the experiment, following.

3.2.3 Introduction

In this experiment, we are presenting two methods that will take benefits from the mappings of the previous section to enrich or extend other resources. The first method will aggregate information to a Website using Schema.org tags as microdata. The second case will extend any of the vocabularies from LOV with properties from Schema.org.

For the first use case, the information will be first obtained from Web Data Commons[60], (Mühleisen & Bizer, 2012). This is a project that extracts data from webs with Microdata,

[60] http://webdatacommons.org/

Microformats[61] and RDFa, providing statistics. The information provided is extracted from the Common Crawl web corpus[62], which is the biggest and most updated dataset for public use providing downloads in the form of N-Quads[63]. The information is stored in instances whose format can be seen in Figure 3-12. From September 2009 to October 2016, crawls of different sizes have been obtained, this information is shown Figure 3-13.

```
<http://www.freebase.com/id/en/built_to_spill> <http://www.freebase.com/id/music/artist/origin>
<http://www.freebase.com/id/en/boise> <http://socialdischord.blogspot.com/>    .
```

Figure 3-11: Example of N-Quad.

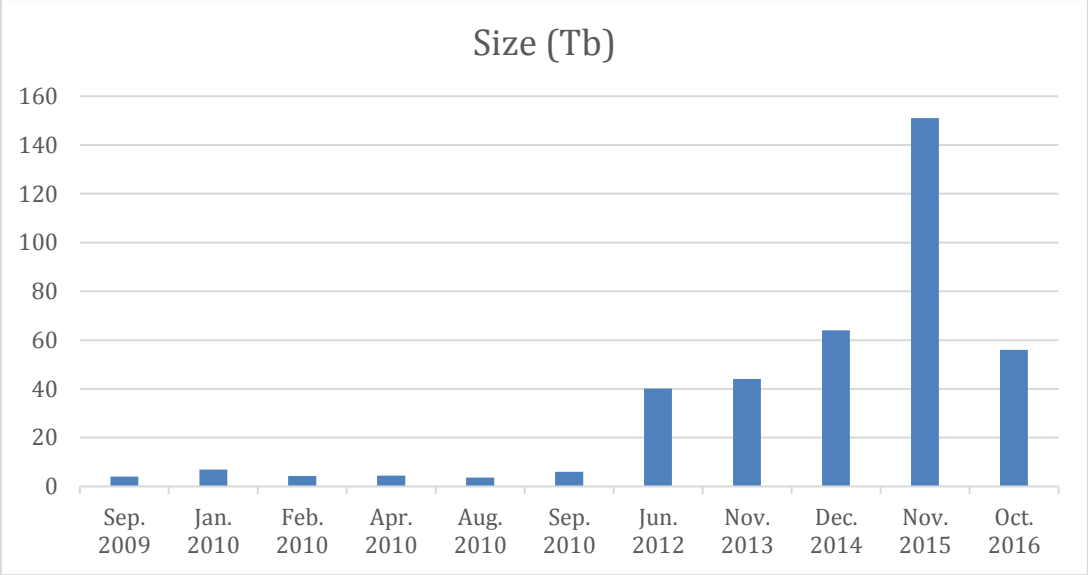

Figure 3-12: Size of the crawls chronologically.

At the moment of the experiment January 2014, the Common Crawl Foundation was providing a dataset of 102 Terabytes with information from about 3.5 billion pages, where over 7.5 billion of N-Quads were found. The data is available in Amazon's Simple Storage Service[64] (S3), which is an interface of Web Services, allowing users to store and retrieve data. This information is freely accessible using the access data Web Service Amazon Elastic Compute Cloud[65] (Amazon EC2), which was designed to make web-scale computing easier for developers.

---

[61] http://microformats.org/

[62] http://commoncrawl.org/

[63] https://www.w3.org/TR/n-quads/

[64] https://aws.amazon.com/es/s3/

[65] https://aws.amazon.com/es/ec2/

Describing the crawl in numbers, it had a size of 44 Terabytes compressed and 148 uncompressed. There are 2,224,829,946 parsed HTML URLs of which, 585,792,337 were URLs with triples. The total number of domains was 12,831,509, having triples 1,779,935 of them. If we count the number of triples in total, it was 17,241,313,916. From all the different formats we can find, we were interested in HTML microdata whose statistics were: a file of 189 Gigabytes with 463,539 domains and 8,795,074,538 triples. In the two following tables, we can see the 5 most used classes and properties from Schema.org.

Class Name	Occurrences in domains
WebPage	69,712
Article	65,930
Blog	64,709
Product	56,388
PostalAddress	52,446

Table 3.15: Most used Schema.org classes according to domains.

Property Name	Occurrences in domains
Article/name	60,340
Blog/name	55,404
Product/name	50,536
PostalAddress/streetAddress	48,358
PostalAddress/addressLocality	47,170

Table 3.16: Most used Schema.org properties according to domains.

The crawl from Web Data Commons will be used to obtain the Websites using Schema.org tags. But we also need some information from the Web of Linked Data that will enrich these Websites, for that purpose we have chosen DBpedia. DBpedia is the biggest dataset and allows users to extract structured information from Wikipedia, (Lehman et al, 2014). At the moment of the experiment, DBpedia could be found in 125 different languages, describing 38,3 million of items, using 3 billion of RDF triples, being 583 million of them in English.

3.2.4 Method

As we have said before, the experiment consists of two different methods that allows us to use the mappings between Schema.org and LOV to enrich a Website or to extend an ontology.

For the first use case, we will start from a particular Webpage, which uses metadata from Schema.org. This tag will be composed of a class and a property from Schema.org and a particular value. If we query to and endpoint of a dataset in the Web of Linked Data with this value, it is possible to obtain extra data to enrich the Website. For example, if we run a query using that particular value against DBpedia, we can find new data related with this value that is not used in the Website and could be aggregated to it. For a better explanation, we will show an example later, but a workflow can be seen in the following Figure.

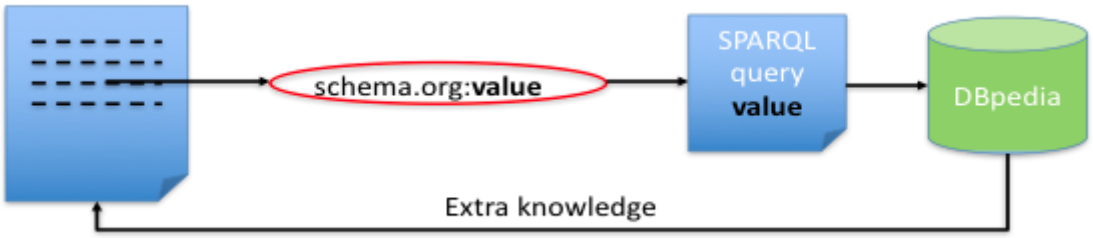

Figure 3-13: Use case for Website enrichment.

The second use case consists of extending a vocabulary from LOV with properties from Schema.org. To achieve it, we need a mapping between two classes that are referring to the same term. For example, Event is a class in Schema.org and also in a LOV vocabulary called Semantic Web Portal Ontology. For both, Schema.org and the vocabulary, the term is describing the same thing, so the properties of Event in Schema.org, can be added to the properties of the same class in the vocabulary. An example will also be given in order to clarify it.

As we have explained before, part of this experiment will be based on a crawl. In particular, this will be used in the first use case of this section. The crawl we have obtained contains all kind of microdata, but we are only interested in the one using Schema.org that we have to filter from the rest. In order to obtain only this kind of data, we have developed a script with Pig Apache[66], which is a language aim to design programs to analyse large dataset to be run

[66] https://pig.apache.org/

on a Hadoop Cluster. Hadoop[67] is a framework used on clusters of computers, so users can process distributed datasets.

The script was developed in four steps. The first step is aimed to obtain only the information using Schema.org, for that purpose we are creating a filter using a proper schema. It is necessary to take into account that some Webmaster can make mistakes when writing the Schema.org tags. To avoid that, the script should only retrieve information using the standard format of Schema.org, which is http://***Class***/***property***. The second step, consists of creating a distinct key, combining the class and property from Schema.org and the value that it has in a particular domain. At this point, for each key we will have every different instance, now it is important to count the total instances for each key. Finally, we create a text file with all the information. The document has the following information for each record: the class and property from Schema.org, the exact value that they have and the number of instances that we have found.

The script was run on February 2, 2014, obtaining a file of 380 Gigabytes uncompressed, having more than 750 million of Schema.org instances and their values. A filter by IRI's with different classes and properties has been made, in order to count how many of these values are useful. After this step, the number of instances has been reduced to 7,783 combinations of classes and properties but we have realized the in some cases no information will be retrieved from DBpedia.

For example, there are cases where the value will be a large text as can be seen the combination of the class "Article" and its property "bodyArticle". To avoid this, the values with more than 255 characters have been discarded. Other example that we are avoiding are those that contain a numeric value, as can be seen the "width" of a "Video" or those whose value is encrypted. Based on this, finally we can work with 1,662 instances.

The final step consists of building the query that will be used with DBpedia. The queries will be obtained using the IRI's from Schema.org with that particular values stored in the new file. The queries are of two different types: the first will only use the particular value from the IRI and the second one, the value with the Schema.org class and property. Following can be seen both examples:

[67] http://hadoop.apache.org/

```
(1)     Query using only values:

PREFIX dbpedia:<http://dbpedia.org/resource/>
SELECT * WHERE {
dbpedia:value ?predicate ?object
}

(2)     Query using Schema.org class and value:

PREFIX dbpedia:<http://dbpedia.org/resource/>
PREFIX rdf:<http://wwww.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT * WHERE{
dbpedia:<value> rdf:schemaClass ?object .
FILTER regex(?object, "<schemaClass>", "i")
}
```

Now, we are describing a particular example, we are using the following instance: http://schema.org/LodgingBusiness/Hotel/addressRegion whose value is “Rio de Janeiro”, which appears in the domain http://mamangua.com. First, we are checking if it is true that this Schema.org IRI has this value in the Webpage. For that purpose, we have used the Goole Structured Data Testing Tool[68] that allows users to obtain all the microdata contained in a web, classifying it by elements, types (giving the metadata form used) and properties. In our example, results show that there is an element tag with Schema.org, which pertains to “LodgingBusiness/Hotel” class and with a property called “addressRegion”.

Keeping in mind what was described earlier, using DBpedia endpoint and running a SPARQL query, we can retrieve extra information about “Rio de Janeiro” that can be added to the initial Website.

[68] https://search.google.com/structured-data/testing-tool

The SPAQL query that we have used is:

```
PREFIX dbpedia:<http://dbpedia.org/resource/>
SELECT * WHERE {
dbpedia:Rio_de_Janeiro ?predicate ?object
}
```

The information retrieved from DBpedia using this query is stored in htttp://dbpedia.org/page/Rio_de_Janeiro. Here, we can find information that cannot be found on the web, like the population of the city, the name of the airport or important monuments like “Cristo Redentor”.

For the second use case, the example could be the following. We need to find a mapping of classes between Schema.org and a vocabulary in LOV. The class “City” from Schema.org is part of the vocabulary Semantic Web Portal Ontology (swpo). The only property that this class has in the vocabulary is “inRegion”. As in both places the class is referring the same term, the properties from Schema.org associated to the class “City” can be added to this class in the vocabulary “swpo”. So, we can extend the vocabulary associating terms like “name”, “URL” or “address” to the class “City”.

3.2.5 Discussion and results

Both methods were applied in order to obtain some results. In the first use case, it has been run queries of both types. We have used each value stored in the file extracted from the Web Data Commons, in relation to the Schema.org IRI’s we have filtered it previously. After running a process, we have obtained new information in Dbpedia: 420,324 times for the first type of queries and 3,529,510 for the second one.

For the second use case, extension of ontologies, we have been able to find at least a mapping between Schema.org and each of the LOV vocabularies. This let us to extend 100% of the vocabularies using the method.

3.2.6 Limitations

We have previously highlighted that there are some limitations because the Webmasters could have written Schema.org tags in a wrong format, so we have only used the ones written in the standard way. Also, we have discarded some cases where the value related with the Schema.org IRI is not giving us any information from DBpedia.

3.2.7 Conclusions and outlook

The aim of this experiment was taking benefits from the mappings obtained previously. We have accomplished that by implementing two methods. The first one allows us to aggregate information retrieved from DBpedia to Websites using Schema.org. The second one, extends the properties of a LOV vocabularies with properties from Schema.org.

3.3 Usage of information from the Web of Linked Data to share scientific knowledge

In this section, we are describing step by step the experiment achieved to accomplish with O1.2. “Present a use case where a dataset of the Web of Linked Data is used to guide a retrieval data strategy.”.

3.3.1 Motivation

We have seen that it is possible to retrieve data from the Web of Linked Data and use it to enrich an independent data source. But maybe, it is also possible to use the information of a dataset to design a data retrieval strategy. During the research, we have been talking about sharing knowledge and open data. If we are working in the field of research: we can propose a data retrieval strategy that could be used to share scientific knowledge.

3.3.2 Introduction

Researchers work to make advancements in science with the purpose, most of the time, of making it available to the rest of the world. Most of the scientific projects are funded with public money, so it seems logical to make the results public to the rest of the population. We have access to thousands and thousands of scientific information like books, papers or results, in order to be used by regular users in their works or daily life. Being this amount of information very big, there is a need of making it easily accessible. In order to solve that, some standards have been developed so the information can be stored in heterogeneous formats.

It exists euroCRIS[69], as one of this organization creating these standards, being a professional non-for-profit association of Current Research Information Systems (CRIS) experts. A CRIS is defined as a tool with the aim of giving access and distribute the scientific information. The objective of the organization is about data access, exchange mechanisms, guidelines or standards over scientific datasets or open institutional repositories. CERIF was

[69] http://www.eurocris.org/

created on that purpose of having a format to interchange data. CERIF allows the user to describe research organizations, which are the relations between them and the results obtained after their scientific works. The model was published as an EC Recommendation to European Member States and a common research model.

Similar to CERIF, there is a project from Cornell University called VIVO[70], (Krafft et al, 2010). VIVO is another dataset which is part of the Web of Linked Data. It has a similar aim of storing the scientific knowledge and sharing information between research organizations. The project consists of an open application based on Semantic Web enabling the discovery of research information across institutions. The starting point is an ontology that organizations use to create local instances where they store their research activities and final results. Then, the different instances of the institutions share information enabling the discovery, networking and collaboration with the data and works from the researchers.

In this section, we will talk about a tool called agVIVO aimed to work with three different sources of research knowledge with two different formats. The first data source will be a VIVO instance that will have a data retrieval strategy obtaining information automatically. This information that will be added will be retrieved from Google Scholar[71], a web search engine that indexes scholar literature. The terms used in the Google Scholar searches will be published works on agriculture. This papers/work will be obtained from the third data source, OpenAGRIS. We have to take into account that OpenAGRIS is an RDF version of AGRIS, which is part of the Web of Linked Data. The information obtained after using the tool could be found into two different standards, VIVO and CERIF.

3.3.3 Materials

As we have said in the previous subsection, in this experiment we are working with three different sources: VIVO, Google Scholar and OpenAGRIS. VIVO and OpenAGRIS are just sources storing information. VIVO is also a dataset in the Web of Linked Data, composed by a network of all VIVO instances that have been created.

If we talk about VIVO, it is an ontology for representing scholarship with an open source software. It was first created by Cornell University and significant partners like

[70] http://vivoweb.org/

[71] https://scholar.google.com

CASRAI[72] (Consortium Advancing Standards in Research Administration Information) or EuroCRIS, the organization responsible of CERIF.

It is able to store, edit, search and browse academic information. Its functioning consists of installing and instance and populate it manually with researchers' information, activities and accomplishments of the institution. Once this is achieved, it enables the discovery of scientific knowledge across the institution and beyond. Content in VIVO is maintained manually or retrieved automatically from local datasets, as the way we are doing in that experiment.

In this experiment, we are using the VIVO instance from Cornell University. If we look at its overall statistics, the instance has: about 26,000 people contributing, 4,400 activities like courses or programs, 29,000 events like competitions or conferences, 12,000 organizations like departments or student groups, 151,000 research items like papers or book chapters and 2,000 topics. A geographic representation of the network that has arisen between organizations collaborating with Cornell in different projects, can be seen in Figure 3-15.

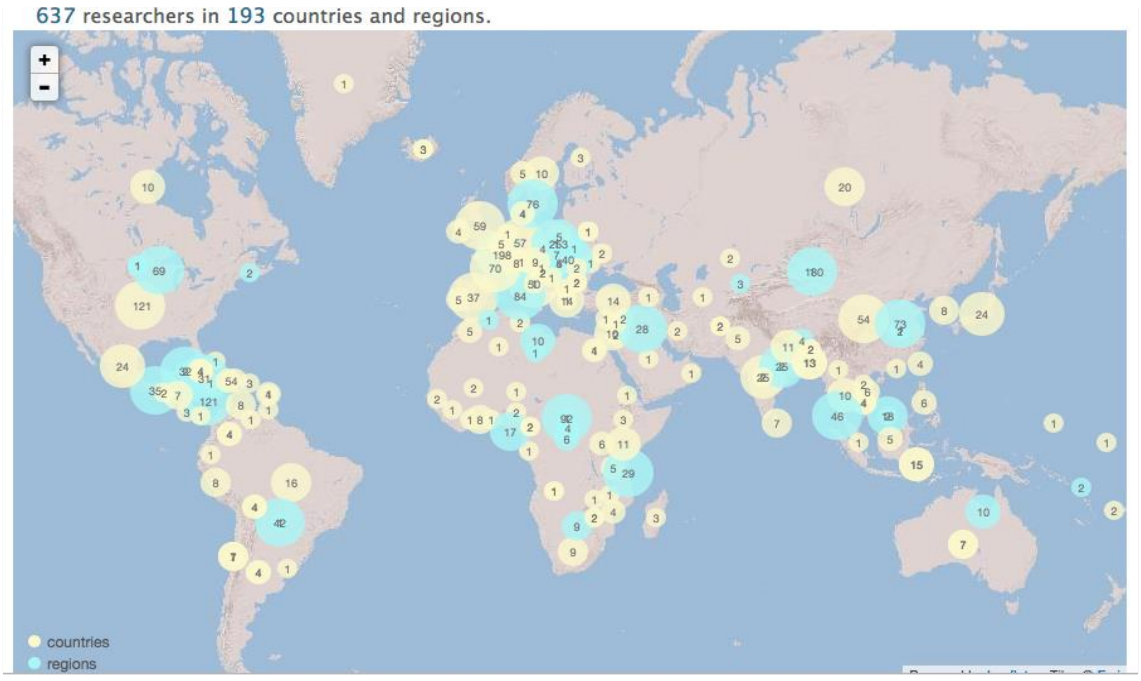

Figure 3-14: Geographical representation of Cornell University VIVO instance[73].

[72] http://www.casrai.org

[73] http://openvivo.org/display/grid.5386.8

When a VIVO instance is populated it can be uploaded to a network of more than 140 institutions in more than 25 countries. This network forms a dataset that is part of the Web of Linked Data. In Figure 3-16, we can see how the different instances are distributed related with the country of origin of the institution. Similar to this but related with the type of institution, we have Figure 3-17.

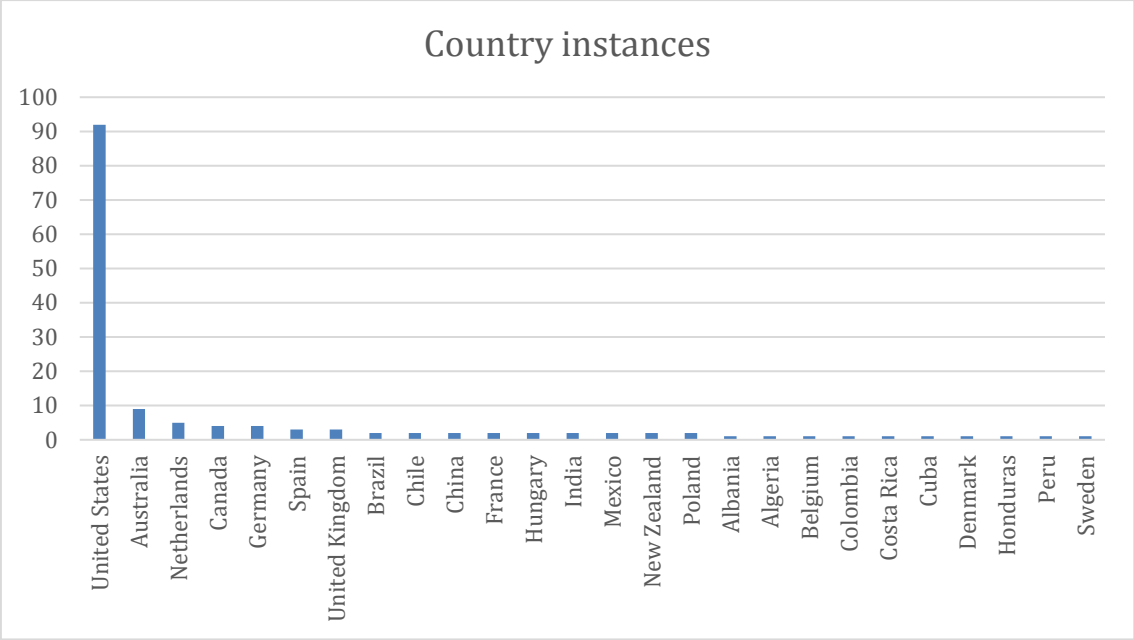

Figure 3-15: Histogram of VIVO instances per country.

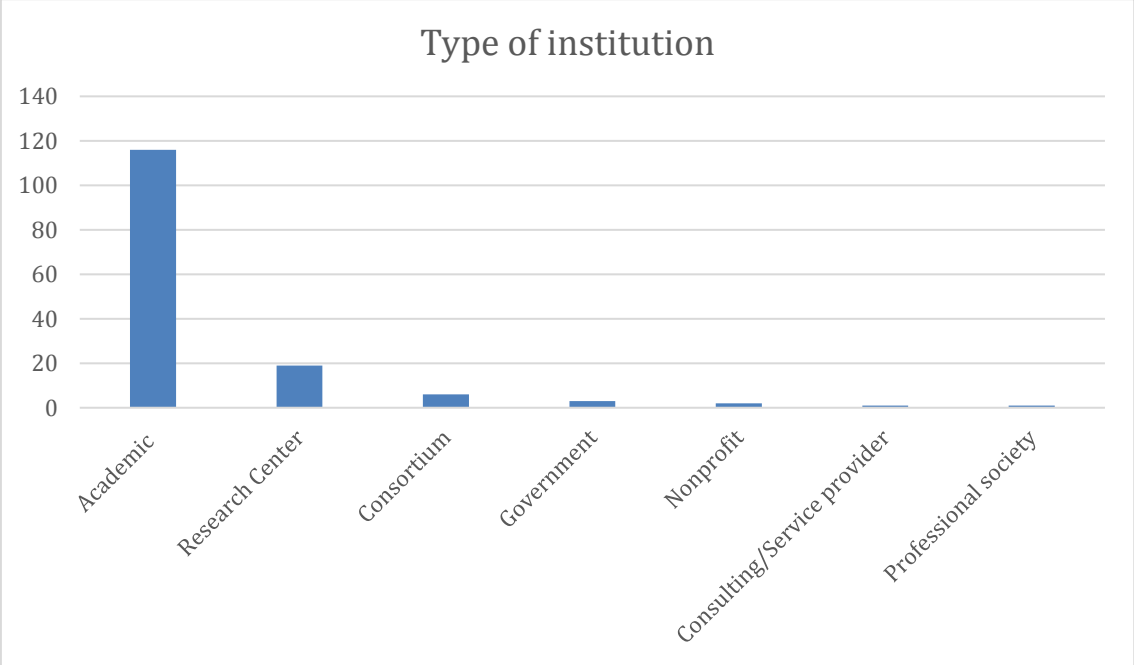

Figure 3-16: Histogram of VIVO instances per institution.

The second data sources we are using is OpenAGRIS, containing a total of 8,977,636 records. It is an RDF representation of AGRIS, a catalogue of scientific references in the field of agriculture. When it shows a paper, it compiles information from various resource, the one named before AGRIS, DBpedia and AGROVOC[74], which is a vocabulary covering all the areas of interest for the Food and Agriculture Organization[75] (FAO). In the following Figures, you can see the information provided by OpenAGRIS. In the first one you can see the main information and in the second one the secondary.

Figure 3-17: Main information provided by OpenAGRIS for a paper[76].

[74] http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus

[75] http://www.fao.org/home/en/

[76] http://agris.fao.org/openagris/

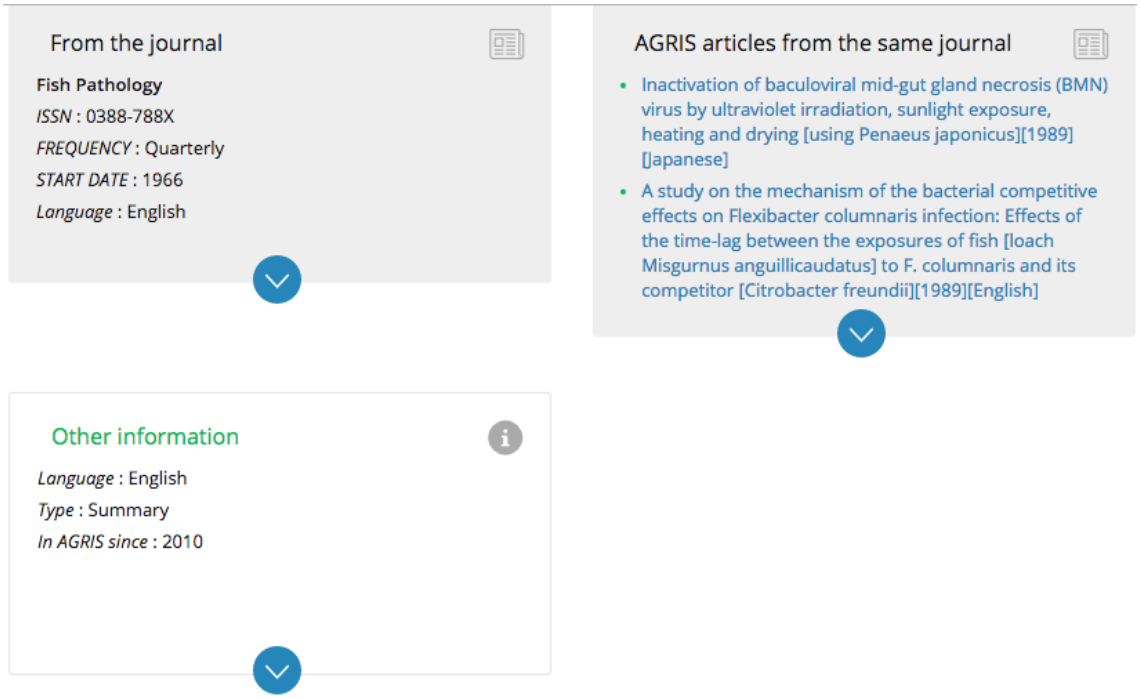

Figure 3-18: Secondary information provided by OpenAGRIS for a paper[77].

The final source of data we are using is Google Scholar. As this source of data cannot be downloaded entirely, we have made a scrapper to obtain some information extra information for a set of OpenAGRIS papers. In total, we have downloaded information of 101 papers with data like: number of citations, references or the full text. The downloaded information has been stored in a database, so our tool could manage it.

3.3.4 Methods

The main objective of this experiment consists of developing agVIVO, then we will use it in an agricultural use case to demonstrate its usage. The architecture of agVIVO, which can be seen in the following Figure, is divided into two modules. First, we have a module called VIVO-io, which will aggregate to a VIVO instance data from Google Scholar papers. To search this information, we are using the titles of the papers stored in OpenAGRIS. The second module is called CERIF2VIVO that lets us to translate the VIVO ontology, once it has been populated, into an instance of a European standard like CERIF.

[77] http://agris.fao.org/openagris/

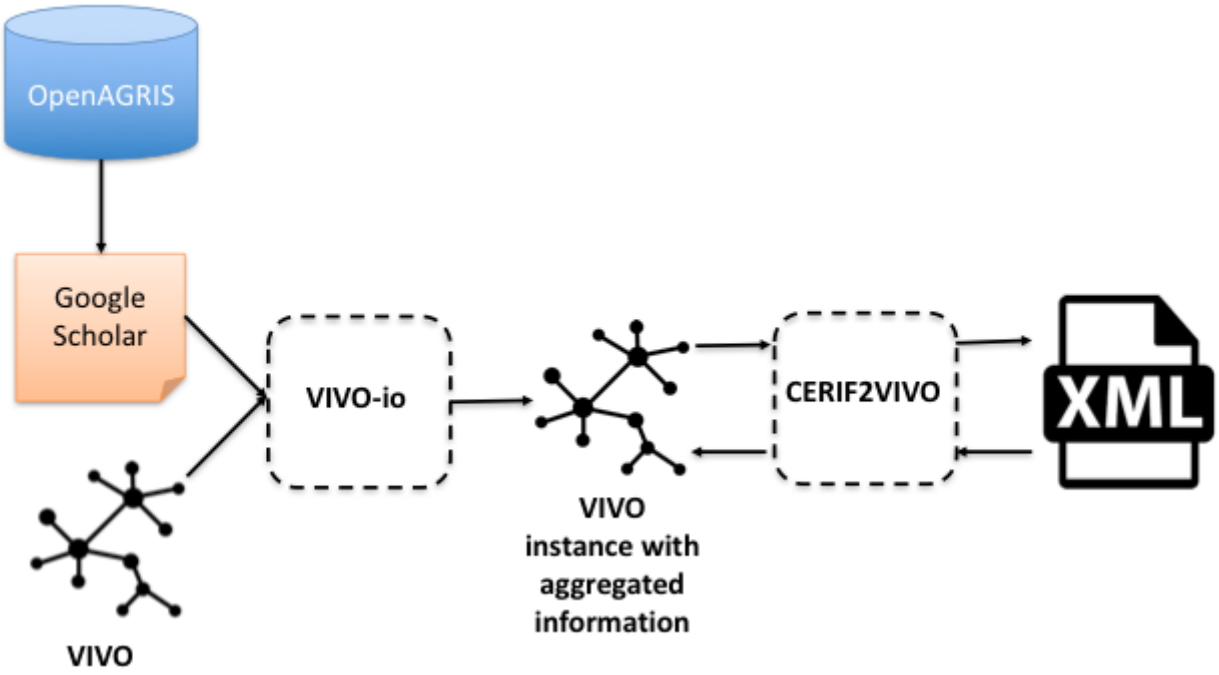

Figure 3-19: agVIVO architecture.

The first module of the tool consists of aggregating data to a VIVO instance automatically. To accomplish that, a Java module has been created using Apache Jena[78], a framework to manipulate ontologies. The best advantage of using the VIVO ontology is that the information can be added automatically, once we have a data source where to retrieve it.

This independent source with information, will be Google Scholar as is one of the biggest catalogues of scientific information. But the problem now is that we need a list of papers to be searched in Google Scholar whose information will be then added to the VIVO instance. For that purpose, we will use OpenAGRIS. OpenAGRIS is a set of agriculture references with more than 7 millions of instances. Using the titles of OpenAGRIS we can search in Google Scholar for new information like full texts or references. Then, we will aggregate this information to our VIVO instance.

Now, we are describing the workflow used to add the information. First, we are using titles from OpenAGRIS to search for new information in Google Scholar, this information will be stored in a database. As the title of OpenAGRIS are stored in AGRIS, a dataset of the Web of Linked Data, we can consider that this dataset has been used to guide the data retrieval strategy. Then we are querying the titles of our VIVO instance with the database. If one occurrence is retrieved, that means that there is new information to be added. For example,

[78] http://jena.apache.org/index.html

using the property “cites” a paper is related with its references. Once we have added the new information from Google Scholar, we have a VIVO instance with aggregated information.

The second module allows us to translate a VIVO instance into CERIF and vice versa. When using VIVO, it arises the problem that its usage is not very extended, as it is not considered a standard and it is not recommended by any public organization. However, we have CERIF, which we have mentioned, is a standard and a recommendation of the EU community, being used for several years. As CERIF is maintained by an organization like euroCRIS and allows to develop a CRIS, translating VIVO to CERIF will be an advantage. To make translations between different formats we need to stablish some mappings, in this case we are using those provided by (Lezcano et al, 2012) and (Lezcano et al, 2013). In the following tables, there are first examples of mappings for the principle terms and then examples of mappings between properties.

CERIF Table	VIVO Class
cfPers	foaf:Person
cfResPubl	bibo:Document
cfResPat	bibo:Patent
cfResProd	vivo:CaseStudy vivo:Dataset
cfFacil	vivo:Facility
cfSrv	vivo:Service

Table 3.17: Examples of mappings between principle terms of CERIF and VIVO.

CERIF Table.Attribute	VIVO Class:property
cfProj.cfURI	Project:webpage only vivo:URLLink
cfProj.cfAcro	Project:description only Literal
cfProj.cfStartDate	Project:dateTimeInterval only DateTimeInterval
cfProj.cfEndDate	Project:dateTimeInterval only DateTimeInterval

Table 3.18: Examples of mappings between properties of CERIF and VIVO.

VIVO is represented with a set of classes and properties in RDF and CERIF uses XML. Using techniques of eXtensible Stylesheet Language Transformation (XSLT), which allows to transform XML documents in other formats, will be a proper way to transform from one to another format. For that purpose, we have developed to stylesheets: first one transforms VIVO into CERIF and the second one does the reverse transformation, (Nogales et al, 2014). Apart from the stylesheets a processor to do the transformations is needed, in our case it will be Saxon[79].

3.3.5 Discussion of the results

As the information obtained from Google Scholar is small, we have tested our tool in one use case just to demonstrate that it accomplishes with the main aim we have proposed. This was aggregating scientific knowledge to a dataset in the Web of Linked Dataset, using another dataset as a way to guide the data retrieval strategy.

Using agVIVO we have been able to add information to a VIVO instance and translate it to CERIF. Taking the Cornell University instance of VIVO which is also part of the Web of Linked Data, we have found that the paper “Pathogenic microorganisms of concern to the dairy industry” written by Kathryn Jean Boor has no references. We have used the following SPARQL against the instance to check it.

```
PREFIX vivo:<http://vivoweb.org/ontology/core#>
SELECT ?subject {
WHERE {?subject vivo:title ?title .
FILTER (REGEX(STR(?title), "title", "i"))
}
```

This paper is also included in OpenAGRIS, so by searching in Google Scholar we can obtain information that is not in the VIVO instance and aggregate it. Figure 3-21 shows all the information that Google Scholar provides from this paper. By querying our database storing the Google Scholar information, we can obtain the new data.

[79] http://saxon.sourceforge.net


[CITATION] **Pathogenic microorganisms of concern to the dairy industry**
KJ Boor - **Dairy**, food and environmental sanitation: a publication ..., 1997 - agris.fao.org
Dairy, food and environmental sanitation : a publication of the International Association of Milk, Food and Environmental Sanitarians (Nov 1997). **Pathogenic microorganisms** of **concern** to the **dairy industry**. Boor, KJ (Cornell University, Ithaca, NY.). Date of publication, Nov 1997. ...
Cited by 22 Related articles All 3 versions Cite Save More

Figure 3-20: Google Scholar snippet.

Using our VIVO-io module, we have been able to add the reference of the paper to the VIVO instance having it more information. As said before the reference have been obtained from the database we obtained scrapping Google Scholar.

Once we have aggregated new information to the VIVO instance, we have translated it into CERIF a standard from Europe not used in the US. This is useful as some institutions managing CERIF CRISs will be interested in working with the information from Cornell University and other institutions using VIVO.

3.3.6 Limitations

The first problem that we have found is about the number of papers that we could retrieve from Google Scholar. As Google Scholar does not provide a free dump, when using the scrapper for a few minutes, this is rejected by the server. So finally, we could only obtain extra information from 101 OpenAGRIS' papers, which is a very little amount.

The other problem comes from previous research when defining the mappings between CERIF and VIVO. As not all the entities from a data source has its equivalent, some of them cannot be translate from one format to another.

3.3.7 Conclusions and outlook

In the development of this experiment, a tool aimed to enrich a particular dataset with information from others has been developed. The development can automatically add new information to a VIVO instance. Then, this instance can be transform into a different format as CERIF and vice versa. During the experimentation, we have been able to combine three different data sources: VIVO which is part of the Web of Linked Data, OpenAGRIS which is the RDF version of another dataset of the Web of Linked Data and Google Scholar.

3.4 Usage of vocabularies in the Web of Linked Data

In this section, we are describing step by step the experiment achieved to accomplish with O2 "Make an analysis of the structure formed by the vocabularies used in the Web of Linked Data".

3.4.1 Motivation

Before creating a dataset, we need to define the different terms we are using. For example, in a dataset of geography in Spain, we need to define the term “city” to create an instance of “Madrid” or “river” for “Guadalquivir”. The part of a dataset responsible of describing the classes and properties of a dataset will be the vocabularies. In the case of the Web of Linked Data, there is a catalogue that registers all the vocabularies that it uses. We have mentioned it before as LOV. If we want to use the information stored in the Web of Linked Data, it will be interesting to know which are the most popular ones. It is also worthy to understand how they are related, so we could understand how they are used. If we know the usage of the vocabularies in the Web of Linked Data, it could be applied to data retrieval strategies.

3.4.2 Introduction

In previous section, we have talked about the Web of Linked Data as a network structure formed by open big data sources. Given the open and democratic nature of the use of vocabularies in Linked Data, having an understanding how the different institutions and communities are using them, is critical for deciding on their potential.

Related with the usage of vocabularies, we have also talked about LOV. LOV initiative is aimed at providing an easy access to the vocabularies used in the Web of Linked Data. It also gives information about how they are related between them by using a vocabulary called VOAF. Finally, it also provides statistics of their use in the Linked Data Cloud.

This information is really useful, but we need to have a general view of the structure formed by vocabularies’ relationships.

By doing this experiment, we are reporting an analysis of LOV vocabularies, metrics about their general characteristics, their relations and how they are use in the Linked Data Cloud.

3.4.3 Materials

In this experiment, we have used a dump of LOV like we have done previously. In this case the dump was downloaded in December 2014 and contained 441 vocabularies.

The other dataset used in the experiment, is a dataset about the usage of the vocabularies in the Web of Linked Data. The crawl, (Schmachtenberg et al, 2014), is a .nq file of 42.68 Gigabytes uncompressed and it contains 188,440,372 N-Quads of 1,014 datasets. The crawl was obtained by Mannheim University in April 2014 using LDSpider[80], (Isele et al ,2010).

[80] http://wiki.planet-data.eu/web/LDSpider

LDSpider works by providing seed URIs, in this case were three sources: datasets from the datahub.io catalogue, URIs from the Billion Triple Challenge and datasets from the mailing list lod@w3.org.

3.4.4 Methods

The experiment is divided into two big analysis: first one will use the dump provided by LOV and the second one the crawl with information from the Web of Linked Data.

For the first analysis, we have developed an iPython notebook that will accomplish two different objectives: the first will be a general analysis of the vocabularies' characteristics and the second a structural analysis by applying SNA techniques. For the first part, we have used the basic libraries like pandas or NumPy to manage the data and Matplotlib to draw graphics. As the dump has RDF information, we need the RDFLib[81] package to obtain the information given by VOAF vocabulary. The SNA analysis has been made with a package called NetworkX[82], which was developed to perform studies of complex networks.

As we have said, the first part of this analysis has been done to obtain the main characteristics of each vocabulary. These characteristics that we are interested in, are characteristics that the vocabularies have in common and are provided by VOAF vocabulary. VOAF gives to each vocabulary has a property called "language" denoting which language is used by the vocabulary. It also exists "classNumber", which gives the number of classes a vocabulary has. Similar to this but related with properties, we can find "propertyNumber". Finally, there is a property called "hasPart" which relates a vocabulary with a Vocabulary Space. A Vocabulary Space is used to know the scope of the vocabulary.

The second part of the analysis consists of obtaining some of SNA metrics, such as diameter, density, clustering coefficient and number of connected components. Also, it is interesting to understand if the network is heterogeneous or homogeneous. As these metrics have been defined before, we will only remark their utility. About structural analysis, there are also some subgraphs formed by VOAF properties that define the relations between vocabularies. These properties are "reliesOn", expressing that a vocabulary extends some classes or properties of another vocabulary. The "metadataVoc" property expresses if a vocabulary uses another one at vocabulary or element level. "usedBy" indicates that one vocabulary uses parts of

[81] https://github.com/RDFLib

[82] https://networkx.github.io/

another vocabulary. The property “extends” denotes that the first vocabulary extends the expressivity of the second vocabulary. The property “specializes” denotes that a vocabulary redefines subclasses or properties of another vocabulary. Property “generalizes” is used in the same way as ”specializes”, but generalising subclasses and subproperties. Another property “hasEquivalencesWith”, indicates that two vocabularies have some equivalent classes and properties. The use of “hasDisjunctionsWith” has the same purpose, but for disjunction classes and properties. Lastly, “similar” is used when vocabularies are similar in scope or objectives. For each property, the analysis has calculated number of nodes, degree centrality, closeness centrality, betweenness centrality, connectivity of the graph and number of connected components.

A crawl that gives us an idea of the state of the Web of Linked Data, will be used in the second analysis of the experiment. The information obtained will say which are the datasets containing most vocabularies, which are the most used vocabularies in the datasets and a comparison between our analysis and previous ones, to measure the precision of ours.

3.4.5 Discussion and results

The first property that has been analysed is that related with the language. In total 44 different languages have been used. If we want to know which are the most used and the distribution of this parameter, Table 3-19 and Figure 3-22 provides this information. Also in Table 3-20 there is information about the largest number of languages used by vocabulary. By analysing these results, we have realized that 41 of the languages are supposed not to be described by a language.

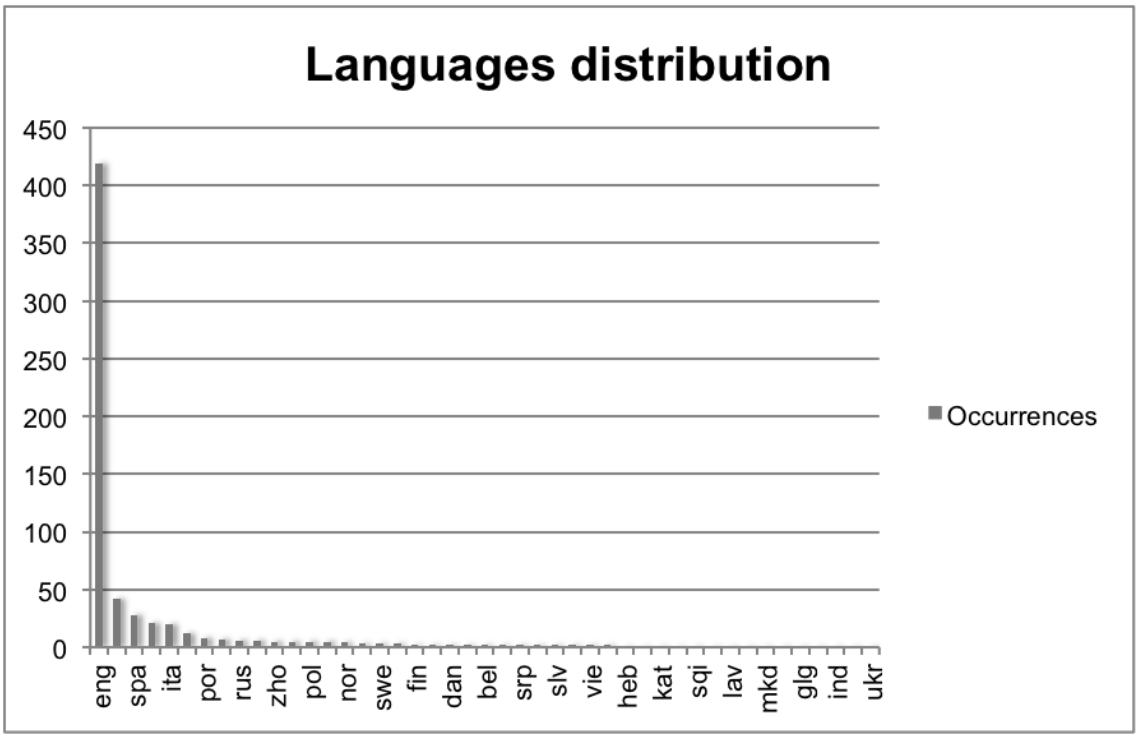

Figure 3-21: Distribution of languages per vocabulary.

Languages	Vocabularies used	Vocabularies used
English	390	88.63%
French	34	7.72%
Spanish	22	5%
German	19	4.77%
Italian	19	4.31%

Table 3.19: Number of vocabularies per language.

Vocabulary	Languages
lgdo	41
mil	17
lingvo	15
bevon	8
geop	8

Table 3.20: Top vocabularies by used languages.

The next properties that have been analysed are these related with the number of classes and properties of vocabulary. This information is shown in the two following Tables and their distributions in the other two Figures.

Vocabulary	Classes
dicom	1592
acm	1469
sio	1414
lgdo	1202
dogont	763

Table 3.21: Top vocabularies by number of classes.

Vocabulary	Properties
dicom	7033
schema	803
rdarel	508
rdag1	455
ebucore	299

Table 3.22: Top vocabularies by number of properties.

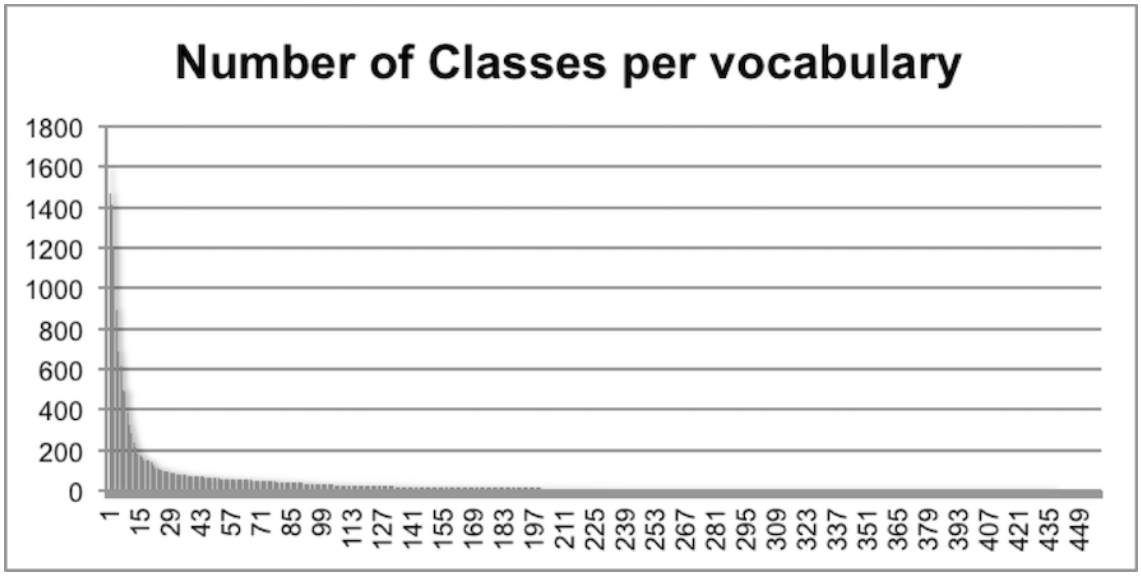

Figure 3-22: Distribution of classes per vocabulary.

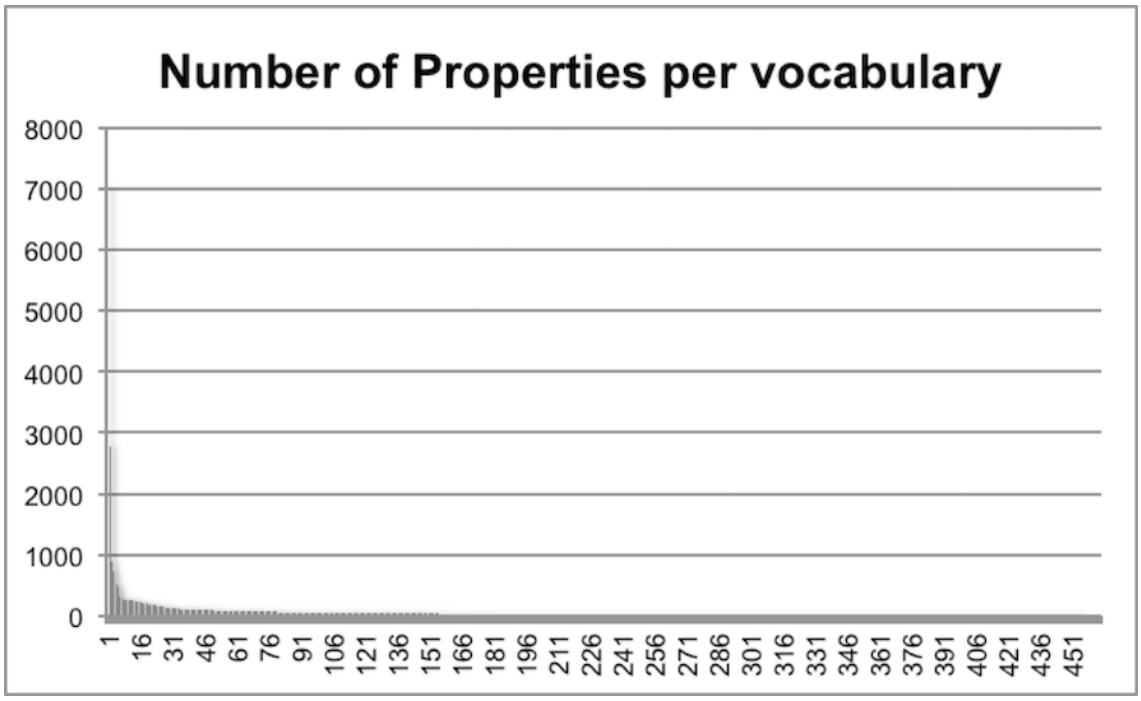

Figure 3-23: Distribution of properties per vocabulary.

The last property to be analysed is that related with Vocabulary Spaces or in other words the scope of the vocabulary. Also, for this property we have a table with the 5 most used Vocabulary Spaces.

Vocabulary Space	Number of vocabularies	Percentage
STATS	30	6.8%
DOC	28	6.3%
API	26	5.8%
META	24	5.4%
QUALITY	23	5.2%

Table 3.23: Vocabulary Spaces with more vocabularies.

For the second part of the experiment, we are doing a SNA by using NetworkX package. But first, we need to see LOV as a graph where the nodes are the vocabularies and the VOAF properties are the edges connecting them. The metrics that have been obtained are: diameter, density, average clustering coefficient, the number of connected components and the type of network. The information is summarized in the following Table.

Metric	Value
Diameter	9 nodes
Density	0.017
Clustering coefficient	0.52
Number of connected components	1st component: 439 nodes 2nd component: 2 nodes
Type of network	Heterogeneous

Table 3.24: SNA metrics of LOV structure.

Another interesting information that we can obtain from the LOV structure consists of making SNA to the subgraphs formed by every VOAF property. This information is also summarized in Table 3.25

VOAF property	Nodes	Degree centrality	Closeness centrality	Connected graph	Connected components
reliesOn	25	0.0416	0.0	False	1 of 9 nodes 4 of 6 nodes 6 of 2 nodes
metadataVoc	396	0.0075	0.0075	True	1 of 390 nodes 1 of 6 nodes
extends	288	0.034	0.046	True	1 of 213 nodes 1 of 12 nodes 1 of 7 nodes 1 of 5 nodes 1 of 4 nodes 5 of 3 nodes 16 of 2 nodes
specializes	306	0.0065	0.0131	True	1 of 271 nodes 1 of 5 nodes 2 of 4 nodes 2 of 3 nodes 8 of 2 nodes
generalizes	56	0.0727	0.0	False	1 of 10 nodes 1 of 5 nodes 2 of 4 nodes 2 of 3 nodes 8 of 2 nodes
hasEquivalences	118	0.0085	0.0	False	1 of 5 nodes 1 of 3 nodes 5 of 2 nodes
hasDisjunctionsWith	20	0.21	0.0	False	1 of 5 nodes 1 of 3 nodes 6 of 2 nodes
similar	18	0.0085	0.0	False	1 of 5 nodes 1 of 3 nodes 5 of 2 nodes

Table 3.25: Analysis of the VOAF properties of relations between vocabularies.

Giving a look to the table above, only the properties “metadataVoc”, “extends” and “generalizes” can help us to obtain some conclusions. These properties are relating a big amount of vocabularies. In these examples, centrality measures are close to zero. That means several for example: in the case of degree centrality, we can conclude that most of the vocabularies are not relate to many other vocabularies. In the case of closeness centrality, the proximity to zero tells us that nodes do not share many terms.

The second analysis of this experiment has been made with a crawl that represents the Web of Linked Data. The information we are interested in, is how vocabularies are by the different datasets. In the following Tables, we can see the number of vocabularies used by the different datasets and the top five datasets with the highest number of vocabularies.

Vocabulary	Number of occurrences	Percentage
rdf	996	98.22%
rdfs	736	72.58 %
foaf	701	69.13 %
dcterms	568	56.01 %
owl	370	36.48 %
geo	254	25.04 %
sioc	179	17.65 %
mvco	157	15.48 %
skos	143	14.10 %
void	137	13.51 %

Table 3.26: Use of vocabularies in datasets.

Dataset	Number of vocabularies
w3.org	102
southampton.ac.uk	49
b4mad.net	36
mit.edu	33
jones.dk	27

Table 3.27: Top datasets by number of vocabularies.

An analysis of the table above gives significant information about the usage of vocabularies. For example, RDF, RDFS and OWL are vocabularies used to model other vocabularies. Also, SKOS and DCTERMS are standards in the Semantic Web. We can conclude that these vocabularies can be considered as part of the “most popular vocabularies”.

3.4.6 Conclusion and outlook

In the experiment, we have made an exhaustive analysis of the LOV structure and the different vocabularies. First, we have obtained a report of the main characteristics of the vocabularies. Then, a SNA has been performed based on the structure formed by the vocabularies and the VOAF vocabulary. Finally, we have made an analysis of the usage of vocabularies in the Web of Linked Data.

Taking into account the first report, we can conclude the following things. As most of the vocabularies used English for their description, most of the datasets will use this vocabulary for their information. As the number of properties and classes in every vocabulary is not very high, the vocabularies seem not be highly specialized in a field. This issue also affects at the time of building a dataset, as the vocabularies do not have a big amount of terms, several vocabularies will be needed. Related with the specialization of the vocabularies, it is also remarkable as the Vocabulary Spaces are widely distributed there is not a dominant scope.

If we talk about the relations of the vocabularies the biggest amount of vocabularies is related by the properties: “metadataVoc”, “extends” and “specializes”. The information obtained conclude that most of the vocabularies are reusing information from another one. We have also concluded that normally a vocabulary is only related with a few by not many terms.

Finally, based on the results of the usage of vocabularies in the Web of Linked Data. We can conclude that the most used vocabularies are those used to model another vocabulary like RDF, RDFS and OWL. Also, Semantic Web standards like DCTERMS or SKOS.

3.5 On the graph structure of the Web of Linked Data

In this section, we are describing step by step the experiment achieved to accomplish with O3 “Make a structural and quantitative analysis of the Web of Linked Data”.

3.5.1 Motivation

We have worked in use cases of data retrieval from datasets of the Web of Linked Data. But we also know that if we build new strategies they are totally related with its structure.

Let's put an example with a similar problem. In logistic, a company has to distribute its warehouses to optimize money, time, etc. If the clients of this company are distributed radially, the solution will be having a big warehouse in the center. If they are distributed forming a circumference, the solution will be having a few little warehouses distributed between the clients.

In the Web of Linked Data, we can have the same problem. If we know which nodes are more important and how are the connected to the others, data retrieval strategies will be different. To achieve these issues, we have to know: the importance of the different datasets in the network structure, the connected components and the structure of the Web of Linked Data.

3.5.2 Introduction

Linked Data uses data retrieval technologies as HTTP and mechanisms of identification like URI. As we have said before this makes the Web of Linked Data as a space of open and structured data based on the Tim Berners Lee principles.

In 2011 Bizer and Heath published the adoption of Linked Data best practices, which have been used to create datasets. In previous sections, we have talked about the formats of these datasets, which use RDF language and are formed by triples. The information represented by these triples could be URIs, that can be look up by using HTTP, or particular values. A triple is also the responsible of interlinking datasets when the subject and predicate belong to different datasets, is what we call RDF link.

By seeing the Web of Linked Data as datasets linked by RDF links, we can understand it as a graph. Here the datasets will be the nodes and the RDF links, the edges. There are several applications created for navigating the data through the structure of the Web of Linked Data. This structure also will evolve when new datasets are added. Only since its creation in 2007, the Web of Linked Data has evolved from a dozen of datasets to more than a thousand.

In this experiment, we deliver a report on the main findings of an analysis of the graph structure of the Web of Linked Data. This led to important insights helping to innovate in data retrieval techniques or a better understanding of the structure itself. First of all, main metrics of the datasets are obtained. Then, a SNA is used to have a general picture of the structure of the Web of Linked Data.

3.5.3 Materials

For that experiment, we have used the crawl provided by Mannheim University that has been described previously. The information provided is in the structure of n-quads with the format:

subject, predicate, object and dataset. We have transformed the crawl in a .csv file that has been cleaned and normalized. Some URIs are ill-formed using Hexadecimal notation for special characters, so these instances, which are 7.32% of the dataset, have not been taken into account. We have also find n-quads that has links to Websites and not to datasets, so they have also been skipped. Finally, an amount of 166 URI seeds could not crawl any information.

3.5.4 Methods

The analysis of the file has been made with an iPython notebook. First of all, the graph has been created. The information of the nodes has been obtained from a file that contains all the name of the datasets, also provided by Manheim University. The information of the edges has been obtained from the file that we have normalized and cleaned before. Then we have used well-known libraries like NumPy, pandas to manage the data and Matplotlib to draw graphics. We have also used NetworkX to create the graph and get SNA metrics. For the creation of the graph, only the instances that connect two datasets have been taken into account. The final result has been a directed graph.

3.5.5 Results and discussion

Using the graph that we have generated, it is possible to obtain some general measures. First, we have to realized that the Web of Linked Data is a disconnected graph, so some measures like average path length cannot be computed. A summary of these general metrics is shown in the following Table.

Metric	Value
Number of vertices	1,014
Number of edges	4,692
Strongly connected	False
Weakly connected	False
Diameter	9
Degree centrality	0.0019
Closeness centrality	0.12

Table 3.28: General statistics of the Web of Linked Data.

Another interesting analysis consists of obtaining the distribution of the information in the Web of Linked Data. As the structure is formed by datasets, there is an interest in knowing which of them are the biggest. In the following Table, you can see the top 5 biggest datasets with its name, number of occurrences and total percentage regarding the Web of Linked Data.

Dataset	Number of occurrences	Percentage
opendata.euskadi.net	81,162,382	43.07%
fr.dbepedia.org	13,767,913	7.3%
dbpedia.org	8,130,084	4.31%
dbtropes.org	6,930,857	3.67%
estatrwrap.ontologycentral.com	5,665,528	3.006%

Table 3.29: Number of occurrences in datasets.

Another metric that interests us is the degree of the nodes. Taking the graph as a whole, the distribution of the degrees, is what we call degree distribution. We are studying the in-degree which is the amount of edges arriving at a node and the out-degree which is the amount of edges leaving a node. Figure 3-26 shows the distribution of in-degree and Figure 3-27 the same for out-degree. In Table 3.30 there are the top 5 datasets regarding its in-degree and the same for out-degree in Table 3.31.

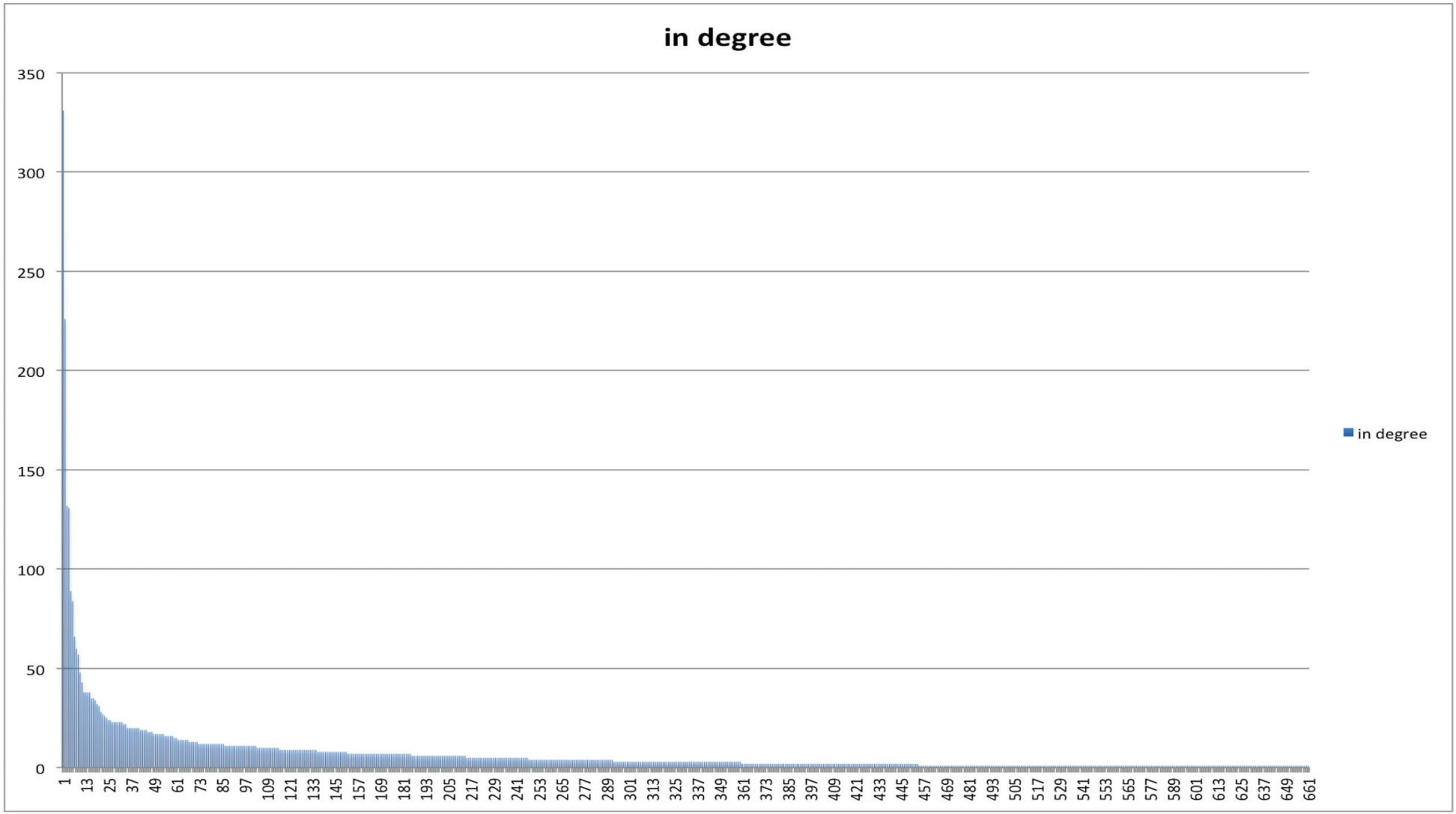

Figure 3-24: In degree distribution.

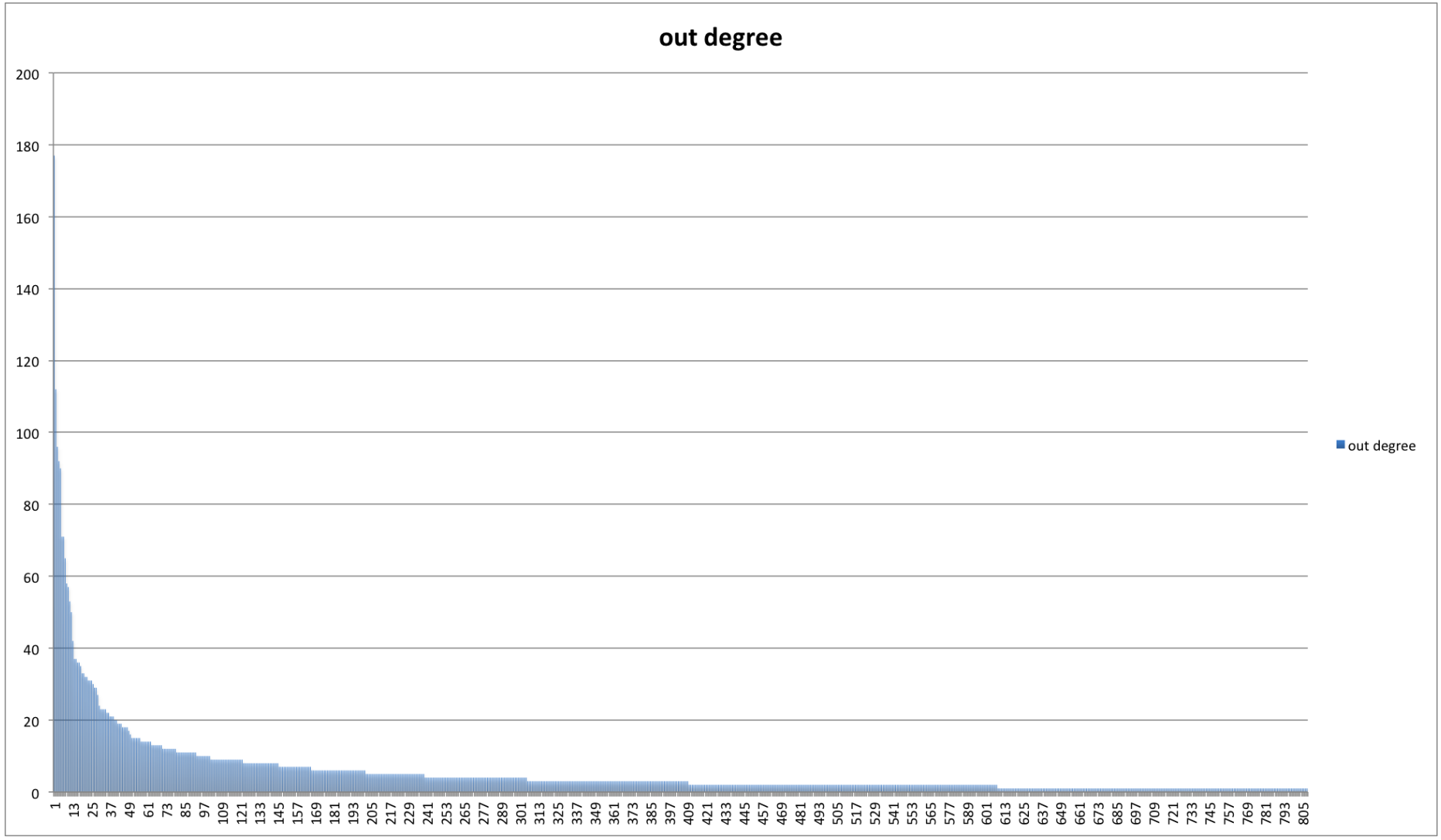

Figure 3-25: Out degree distribution

In-degree	Dataset
331	w3.org
226	dbpedia.org
132	reference.data.gov.uk
131	geonames.org
89	semalink.net

Table 3.30: Top in-degree datasets.

Out-degree	Dataset
117	dbpedia.org
112	semanticweb.org
96	data.semanticweb.org
92	bibsonomy.org
90	fr.dbpedia.org

Table 3.31: Top out-degree datasets.

Another interesting part of SNA is studying the connectivity. This metric allows us to define the structure of the graph; in which sets the nodes are grouped and how they are connected between them. There are two types of connected components: strongly and weakly. After our analysis, the strongly connected components have: one of 511 nodes, another of 3, one of 8, three formed by a pair of nodes and the remaining 486 of only one. If we talk about weakly connected components, we have: one of 904 nodes and 110 of one node.

Taking into account the results related with connectivity, it will be important to discover if the graph complies with the bow-tie theory of (Broder et al, 2000). According to the paper, the bow-tie structure has five components:

- SCC, the strongly connected component that is the core of the structure.
- IN, is made by datasets that can reach the SCC component but cannot be reached.
- OUT, similar to the IN component but formed by datasets that are reached from the SCC component.

- TUBES, has nodes that are not in the SCC component, are reachable from IN and can reach OUT.
- TENDRILS, are datasets that cannot reach and are not reachable from SCC, but belong to IN or OUT components.
- DISCONNECTED, datasets that has no connections. It cannot be considered as a real component of the structure.

In Table 3.32 is the information about the number of nodes of each component.

Components	Nodes
SCC	511
IN	283
OUT	101
TUBES	0
TENDRILS	9
DISCONNECTED	110

Table 3.32: Bow-tie components.

Finally, you can see a graphic of the structure in the following Figure. Here the SCC is coloured in yellow, IN in green, OUT in red, TENDRILS in purple and DISCONNECTED in blue., Nogales et al (2018).

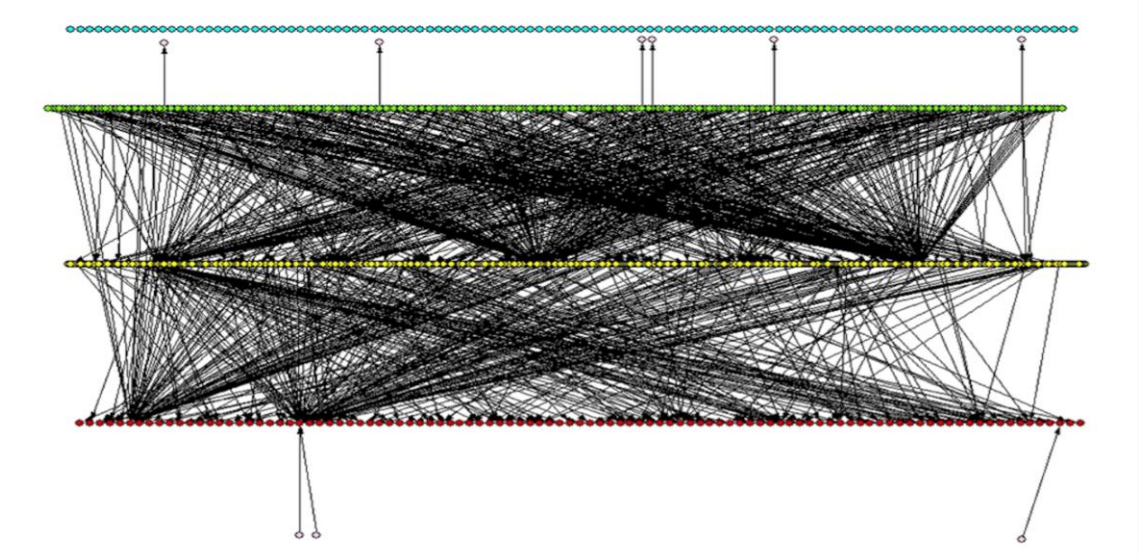

Figure 3-26: Bow-tie structure.

3.5.6 Limitations

The main limitation of this experiment is given by the dump provided by Mannheim University. As we have said before 166 of the datasets that were supposed to be part of the Web of Linked Data could not be crawled. That means that they were part of the URI seeds used with LDSpider but finally no information about them was found. Another problem that we have found is that related with the ill-formed n-quads making difficult to know to which dataset they belong. Finally, some of the n-quads were joining regular Websites, having to discard them also.

3.5.7 Conclusions and outlook

In this experiment, we have tried to give some information about the characteristics and a global view of the structure of the Web of Linked Data. The first metrics like diameter and closeness and degree centrality tell us that the structure is compact and the distance between nodes is low. Also, we have discovered that most of the nodes have a reasonable number of nodes. If we talk about the datasets, we know that Open Data Euskadi is the biggest and WordNet 2.0 and DBpedia the ones with more connections. Finally, we have demonstrated that the structure follows the bow-tie theory.

All the metrics measured in the experiment can be used to perform data retrieval strategies. The dataset tops are useful for rankings. Also, if we make regular studies of the state of the Web of Linked Data, we can understand and predict how it evolves.

4. CONCLUSIONS

In previous sections, the motivation and objectives of this research have been described. The experiments carried out to reach the objectives have also been depicted with detail. In this section, we are putting in common all these elements: relating the objectives with the experiments and its results.

The section is divided into the following subsections:

- First, a table relating the objectives with our contributions is provided. This allows readers to understand how each objective has been accomplished.
- Once the objectives have been accomplished and the questions have been answered, we will expose the contribution.
- Finally, general conclusions of the research have been described and future works have been proposed.

4.1 Attainment of objectives

O1. Make a structural and quantitative analysis of the Web of Linked Data	
O1.1. Make an analysis of the overall structure.	A SNA of the structure formed by the different datasets has been made. We have obtained metrics like diameter, centralities and degrees. By knowing general characteristics of the structure, we can understand the behaviour of the data retrieval strategies.
O1.2. Make an analysis of the most important datasets.	By making a SNA, we can also discover which are the biggest datasets and which have more connections with the rest of the graph. This application of SNA has been used in data retrieval strategies. By applying this information, we can know where to start the searches or which are the key datasets to go through all the structure.
O1.3. Check if the structure of the Web of Linked Data accomplishes with the theory of the bow-tie.	The study of the bow-tie theory has been applied in other researches when studying the Web of documents. It also can be used in data retrieval strategies as it groups the datasets into different components based on how they are interconnected between them.

Table 4.1: Attainments of Objective 1.

O2. Make an analysis of the vocabularies used in the Web of Linked Data.	
O2.1. Make a quantitative report of the characteristics that all the vocabularies have in common.	All the vocabularies have some common characteristics that are described by a vocabulary called VOAF. By applying statistical techniques, we are obtaining a report that will tell us which languages are used, how many terms have a vocabulary and the scope. This information can be used to obtain a general picture of the vocabularies.
O2.2. Understand the structure formed by the relations of the different vocabularies.	Besides VOAF properties describing general characteristics of the vocabularies, there are properties that describe how the vocabularies are related between them. If we make a SNA for these properties, we can obtain a report about how the terms are used between vocabularies. Another part of the SNA could be, obtaining general metrics of graph theory. This information is very useful when developing data retrieval strategies and to understand how the vocabularies behave.

O2.3. Report the usage of the different vocabularies in the datasets of the Web of Linked Data.	Vocabularies are responsible of describing the information that are represented in a dataset. By using statistical techniques, we can obtain the usage of the vocabularies. This also helps us to understand the behaviour of the datasets and the importance of the different roles that the vocabularies have, for example: vocabularies to model other vocabularies or those recognized as standards.

Table 4.2: Attainments of Objective 2.

O3. Develop new strategies to manage data from the Web of Linked Data.	
O3.1. Present a use case in which information from the Web of Linked Data is aggregated to an independent resource.	As the main objective of the research is to make a study of the Web of Linked Data and its components, in order to have better data retrieval strategies. First, we need to demonstrate that, we can retrieve data by developing our own method. We have developed a data retrieval strategy which uses LOV as a link between Schema.org and the Web of Linked Data. By applying ontology mapping techniques, we have been able to aggregate information to retrieve information from a dataset of the Web of Linked Data, DBpedia, and aggregate it to independent resource like Websites using Schema.org. These mappings have also been used to extend ontologies with Schema.org properties. This is a demonstration that new data retrieval strategies can be developed, so there is also an interest in studying the structure of the Web of Linked Data.

O3.2. Present a use case where a dataset of the Web of Linked Data is used to guide a retrieval data strategy.	Another way to take benefits of the information stored in the Web of Linked Data is to use it as a way to guide a data retrieval strategy. In that case we have used OpenAGRIS, which is the RDF version of a repository of scientific papers in agriculture called AGRIS. The titles of the papers stored in OpenAGRIS are used as a guide to search information in another source like Google Scholar, whose information will be added to VIVO another dataset of the Web of Linked Data. Again, building a data retrieval strategy with datasets of the Web of Linked Data tell us about the importance of knowing its structure.

Table 4.3: Attainments of Objective 3.

4.2 Overall contributions

This research contributes to the current state of the art with the following results:

- The demonstration that the information stored in the Web of Linked Data can be used by data retrieval strategies:
  - The first strategy consists of obtaining information from a dataset, so it could be used in an independent data source.
  - The second strategy uses information from a dataset as a guide with the data we want to retrieve from other data sources.
- The data obtained by applying SNA to LOV.
- The data obtained by applying SNA to a crawl representing the Web of Linked Data.
- The numerical data obtained from the two use cases of the first data retrieval strategy, the one that uses Schema.org, can also be relevant for the research.
- The metrics proposed in the paper (Nogales et al, 2017) can also be useful for other researchers that want to obtain more specific metrics from LOV.

Also, the background reviewed in this research can be considered important as it is very complete and actual.

In the next subsections, we will cover the contributions made for each objective proposed at the beginning of the research.

4.2.1 Contributions to make a structural and quantitative analysis of the Web of Linked Data.

As we have said before studying the structure of a data source with SNA techniques is very helpful if we want to make data retrieval faster and more accurate. In other words, if we want to improve data retrieval strategies, having a clear picture about how the information is distributed, and the different data sources connected between them is important.

A clear example of the importance of applying these kinds of techniques or metrics to this kind of structures formed by different data source is (Page et al, 1998). This paper is the starting point of Google search engine by defining a new metric applied to the Web of Documents. In this case, the new metric was more effective than the previous ones, that's why Google became started to emerge as a new search engine between their competitors.

Before this research there were several SNA applied to parts of the Web of Linked Data, projects giving some statistics like vocabulary usage or papers providing some metrics of part of the structure. But if we want to develop new rankings or accurate data retrieval strategies, we need to work with the whole structure.

That's why in this work, we have joined all the elements. We have worked with the most updated crawl of the Web of Linked Data and we have obtained the biggest amount of metrics that we could.

With this study, it has been possible to know: which are the biggest datasets, which are the ones that are more connected to the others, the characteristics of the structure or how they are grouped in sets with different characteristics. The information provided by the study, can be used by researchers interested in working with the stored data. If we know which is the dataset with more connections to the others, we know that this one will have more influence in the structure. Knowing the main characteristics of the structure, we will know if it is necessary a lot of hops to go through the whole structure.

The results tell us that Open Data Euskadi is the biggest dataset or that WordNet 2.0 and DBpedia are those with more connections to the rest of the structure. We also discover that the structure is very compact and normally there is a low distance between every pair of nodes. Also, we know that normally the datasets have a reasonable amount of edges leaving

or reaching them. Finally, the accomplishment of the bow-tie theory allows us to divide the datasets into groups having each it owns characteristics and behaviour.

4.2.2 Contributions to make an analysis of the vocabularies used in the Web of Linked Data.

The structural analysis of the Web of Linked Data is very important. The point is that it is not only formed by the datasets. The datasets contain information about a particular theme: government data of a region of the world or gene products. Then, for having the different instances in these datasets, we need to describe what they are. For that aim, there are vocabularies which describe the different terms we need.

There is a catalogue that tries to compile the vocabularies used in the Web of Linked Data. Each vocabulary has information of some main characteristics that all of them have in common and also about how they relate to each other. This information can also be used in data retrieval strategies with more specialized aims. Maybe a data retrieval strategy could be focused on getting terms in the field of biology, for example.

Before this research, there were studies about the usage of vocabularies in the Web of Linked Data or the analysis of some characteristics in small datasets of vocabularies. In this experiment, we have obtained more metrics and we have worked with a bigger set of vocabularies.

The contributions at this step have been: a clear report about the characteristics of the vocabularies, how they are related between them and its usage in the different datasets of the Web of Linked Data.

We can conclude that most of the information in the Web of Linked Data is in English, the vocabularies are not highly specialized, a few of them are necessary when building a new dataset and there is not a particular field that have more vocabularies than others. By taking into account the usage information, we know that vocabularies used to model other vocabularies and those that are considered standards are the most popular.

Apart of using this information for developing new data retrieval strategies. They could be used in applications trying to optimise the number of vocabularies. Also, when developing datasets, choosing the most completed vocabularies. Finally, it can help to find errors and inconsistences when creating a dump file of the Web of Linked Data.

4.2.3 Contributions to develop new strategies to manage from the Web of Linked Data.

Data retrieval strategies are one of the techniques that take benefits from the application of SNA to a data structure. Having a good data retrieval strategy will give the users more accuracy in results and will give the responses in shorter time. A user can also be interested

in obtaining a special type of information, maybe from a particular field or with special characteristics.

Data retrieval strategies are directly related with the structure of the data source we want to exploit. So, there is a need to demonstrate that we can retrieve information from the Web of Linked Data. The information of a data source can be used in different ways. We can use it as the place where the information is going to be retrieved or as a way to guide our data retrieval strategy. In this research, we have proposed two methods that use datasets from the Web of Linked Data in both ways.

It is known that there have been methods which have retrieved information from the Web of Linked Data before our research. In order to make a stronger work, we have decided to build our owns. First one, is focused in obtaining information from a data source and will take advantage of Schema.org vocabulary, as it is used in Websites and is also a vocabulary that is part of LOV. The second one, will use a data source as a guide to search information in other resource and aggregate it to a data source in the Web of Linked Data.

By building these data retrieval strategies, we have accomplished several issues. First, we have demonstrated that information can be retrieved from the Web of Linked Data and users can build their own methods. Second, we have built a bridge between Schema.org when it is used as a mark-up language in Websites and the Web of Linked Data. Third, we have used this method to enrich Website and to extend ontologies. Finally, we have used the information of a Web of Linked Data source as the guide to query information from other data sources.

4.3 Overall limitations

After developing all the experimentation, we can depict the main limitations of the whole research. The following Table links every limitation with a description of what it affects.

Limitation	Consequence
LOV vocabularies are sometimes not available.	Not all the vocabularies could be used to stablish mappings
Syntactic mappings are not as accurate as they could be.	There are cases of mappings that have not been considered. Multiwords containing symbols like "-", words belonging to British and American English (for example, Organization and Organisation) and synonyms.
Semantic mappings have problems with disambiguation.	This task could not be made automatically.
LODStats has errors in 1185 datasets.	The statistics provided at this point cannot be considered accurate.
Webmasters could have written Schema.org tags with typos.	Only the cases when Schema.org has been used in the standard format has been taken into account.
Scraping information from Google Scholar is not allowed.	Only a few papers from OpenAGRIS can be used for that use case.
CERIF is a standard from the European Union. VIVO is not a standard, is an extended format created in the US.	Some entities could not be convert from one format to the other and vice-versa.
The dataset provided by Mannheim University could not crawl information from 166 datasets.	When making the SNA of the Web of Linked Data, these datasets have not been taking into account.
Some of the n-quads in the Mannheim dataset were ill-formed.	These links couldn't also be taking into, reducing the possible the statistics related with the degree of the nodes.
Some of the n-quads in the Mannheim dataset were linking regular Websites.	These links couldn't be taking into account as part of the structural analysis of the Web of Linked Data.

Table 4.4: Limitations vs Consequences.

4.4 Conclusions and future works

The aim of this research was to make a structural analysis of the Web of Linked Data and its different components. The knowledge of this structure could be used in the future to perform better data retrieval strategies. So, another important thing was to demonstrate that we could develop some strategies that allows users to query the information stored in it.

The first objective consisted on demonstrate that information from the Web of Linked Data could be retrieved and be used in data retrieval strategies. For that purpose, we have design two different strategies. The first one takes advantages of a set of mappings between Schema.org and LOV. These mappings were obtained by developing a script. With this information, we have been able to use it in two different use cases: the first one, consisted of aggregating information from DBpedia to a Website, the second one was used to extend a vocabulary from LOV with properties from Schema.org. The second data retrieval strategy has used a dataset from the Web of Linked Data as part of the strategy design of the strategy. In particular, we have used euroCRIS as the way to guide the search of information in other sources, aggregating it to another dataset. This dataset was the network formed by VIVO instances which is also part of the Web of Linked Data.

Once we demonstrated that there are ways to obtain information from the Web of Linked Data, there is an interest in studying its structure and the structure of its different components. We have first started with a SNA of LOV, which is a catalogue that comprises the vocabularies used in the Web of Linked Data. This analysis of LOV has consisted of obtaining some statistics of main characteristics that of the vocabularies have in common. Also, we have studied the structure formed by them by studying the different relations between terms. Finally, an analysis of the usage of the vocabularies in the datasets of the Web of Linked Data has been made.

The last part of the research was the analysis of the Web of Linked Data as a structure. Here, we have made a SNA, first obtaining some general metrics and them by analysing how the datasets are connected between them.

In future works, mappings between Schema.org and LOV will be used to improve search engines searches. Also, they could be used to create SPARQL-federated queries. As these queries need to retrieve information from different data sources, the mappings could be combined to improve the accuracy in the queries.

The metrics and SNA from LOV, can be used to create better data retrieval strategies. It also has utility in applications trying to optimize the vocabularies depending on the metrics a user is interested in. When working with datasets, the vocabularies' metrics can be used to obtain the more complete datasets that the user could need. Other application is, when a user is creating a new dataset and want to reuse terms from are just created in the Web of Linked Data and which are those that need to be created for the first time. Finally, the information can be used by data curators of the Web of Linked Data when they need to provide dumps that do not contain inconsistences and errors.

The information obtained from the analysis of the Web of Linked Data could also be used in the design of data retrieval strategies. Also, it is very useful to know how the structure evolves during the time and how the datasets behave between them. Another future work will be making a deeper analysis of the components of the bow-tie as they are differentiated by how they are connected, they will be very useful in data retrieval strategies.

REFERENCES

1. Heath, T., Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space (Vol. 1)*. San Rafael, California: Morgan & Claypool.
2. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1-22.
3. Langville, A. N. & Meyer, C. D. (2005). A Survey of Eigenvector Methods of Web Information Retrieval. *The SIAM Review, 47*(1), 135-161.
4. Gregory, K., Groth, P. T., Cousijn, H., Scharnhorst, A. & Wyatt, S. (2017). Searching Data: A Review of Observational Data Retrieval Practices. *CoRR*, abs/1707.06937.
5. Johnsen, L. (2012). HTML5, Microdata and Schema.Org - Towards an Educational Social-semantic Web for the Rest of Us? In *M. Helfert, M. J. Martins & J. Cordeiro (Eds.), CSEDU*, (1), (p./pp. 101-104)
6. Jörg, B. (2010). CERIF: The Common European Research Information Format Model. *Data Science Journal*, *9*, CRIS24-CRIS31.
7. Vandenbussche, P., Atemezing, G., Poveda-Villalón, M. & Vatant, B. (2016). Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web. *Semantic Web Journal, 8*(3), 437-452.
8. Jamali, M., & Abolhassani, H. (2006). Different Aspects of Social Network Analysis. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI '06), *IEEE Computer Society,* (p./pp. 66-72*),* Washington, DC, USA.
9. Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1-7), 107-117.
10. Page, L., Brin, S., Motwani, R. & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. In *Proceedings of the 7th International World Wide Web Conference*, (p./pp.161-172), Brisbane, Australia.
11. Martín, M.S., & Gutiérrez, C. (2006). A Database Perspective of Social Network Analysis Data Processing. In *International Sunbelt Social Network Conference is the official conference of the International Network for Social Network Analysis (INSNA).*
12. Mincer, M.& Niewiadomska-Szynkiewicz, E. (2012). Application of social network analysis to the investigation of interpersonal connections. *Journal of Telecommunications and Information Technology.* 2, 81–89.
13. Adamic, L. A. & Glance, N. (2005). The political blogosphere and the 2004 U.S. election: divided they blog. *Proceedings of the 3rd International Workshop on Link Discovery,* (p./pp. 36-43), New York, NY, USA.

14. Deco, J.D., González, A.M., Diaz, J.I., Mato, V., García-Frank, D., Alvarez-Linera, J., Frank, A., & Hernandez-tamames, J. (2013). Machine learning and social network analysis applied to Alzheimer's disease biomarkers. *Current topics in medicinal chemistry, 13*(5), 652-662.
15. Walther, O. (2015). Social Network Analysis and Informal Trade. *Working Paper No. 01/15, University of Southern Denmark.*
16. Koschade S. (2006). A social network analysis of Jemaah Islamiyah: The applications to counterterrorism and intelligence. *Studies in Conflict & Terrorism, 29*(6), 559-575.
17. Oehler, K., Sheppard, S.C., Benjamin, B. & Dworkin, L.K. (2007). Network analysis and the social impact of cultural arts organizations. *Working Paper, Center for Creative Community Development.*
18. Soares, A.E. & Lopes, M.P. (2014). Social networks and psychological safety: A model of contagion. *Journal of Industrial Engineering and* Management, 7(5), 995-1012.
19. Wang, W., Man, H. & Liu, Y. (2009). A framework for intrusion detection systems by social network analysis methods in ad hoc networks. *Security and Communication Networks, 2*(6), 669-685.
20. Castillejo E., Almeida A., & López-de-Ipiña D. (2012) Social Network Analysis Applied to Recommendation Systems: Alleviating the Cold-User Problem. In *Bravo J., López-de-Ipiña D., Moya F. (eds), Ubiquitous Computing and Ambient Intelligence*, 7656, (p./pp. 306-313), Berlin, Heidelberg.
21. Renoust, B., Ngo, T. D., Le, D.-D. & Satoh, S. (2015). A Social Network Analysis of Face Tracking in News Video. In *K. Yétongnon & A. Dipanda, 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS),* (p./pp. 474-481), Bangkok, Thailand.
22. Papadimitriou, A., Katsaros, D. & Manolopoulos, Y. (2009). Social Network Analysis and Its Applications in Wireless Sensor and Vehicular Networks. In A. B. Sideridis & C. Z. Patrikakis (eds.), *Third International Conference, e-Democracy,* (p./pp. 411-420), Athens, Greece.
23. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. (2000). Graph structure in the Web. *Computer Networks*, 33(1-6), 309-320.
24. Nogales, A., Sicilia, M. A., & Barriocanal, E. G. (2018). On the Graph Structure of the Web of Data. *International Journal on Semantic Web and Information Systems (IJSWIS), 14*(2), 70-85.
25. Saye, J. D. (2001). The organization of electronic resources in the library and information science curriculum. *OCLC Systems & Services*, 17(2), 71-78.
26. Harvey, D. R. & Thompson, D. (2010). Automating the appraisal of digital materials. *Library Hi Tech*, 28(2), 313-322.
27. Bush, V. (1945). As we may think. *Interactions, 3(2),* 35-46.
28. Licklider, J.C.R. (1965). Libraries of the Future. *Cambridge, MA: The MIT Press*.
29. Waters, D.J. (1998). What are digital libraries? *CLIR Issues, 4 July/August.*

30. Leiner, B. M. (1998). The NCSTRL Approach to Open Architecture for the Confederated Digital Library. *D-Lib Magazine, 4.*
31. Arms, W.Y. (2000). *Digital Libraries. MIT Press*, Cambridge, MA.
32. Borgman, C. L. (2000). From Gutenberg to the global information infrastructure: access to information in the networked world. *Education for Information*, 18, 339-356.
33. Smith, A. (2001), *Strategies for Building Digitized Collection*. Washington, D.C. Digital Library Federation, Council on Library and Information Resources.
34. Crow, R. (2002). The Case for Institutional Repositories: A SPARC position paper. *The Scholarly Publishing & Academic Resources Coalition,* 1–37.
35. Koutsomitropoulos, D.A., Tsakou, A.A., Tsolis, D.K., Papatheodorou, T.S. (2004). Towards the development of a general-purpose digital repository. In *Proceedings of the 6th International Conference on Enterprise Information Systems,* 5, (p./pp. 271-278), Porto, Portugal.
36. Hayes, H. (2005). Digital repositories: Helping universities and colleges. *JISC Briefing Paper: Higher Education Sector.*
37. Pappalardo, K.M., Fitzgerald, A.M., Fitzgerald, B.F., Kiel-Chisholm, S.D., O'Brien, D. & Auston, A. (2007). A Guide to Developing Open Access Through Your Digital Repository.
38. Sharif, R.M. & Uhlir, P.F. (2009). An inventory of resources for creating an open institutional repository. *Program on Digital Knowledge Resources and Infrastructure in the Developing Countries.*
39. Bargmeyer, B.E. & Gillman, D.W. (2000). Metadata standards and metadata registries: an overview. In *International Conference on Establishment Surveys II,* Buffalo, New York.
40. Caplan, P. (2003). *Metadata Fundamentals for All Libraries*. Chicago: American Library Association.
41. Guenther, R.S. (2004). Using the Metadata Object Description Schema (MODS) for resource description: guidelines and applications. *Library Hi Tech*, 22(1), 89-98.
42. Cundiff, M.V. (2004). An introduction to the Metadata Encoding and Transmission Standard (METS). *Library Hi Tech*, 22(1), 52-64.
43. Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 284, 34-43.
44. Fensel, D., Domingue, J. and Hendler, J. (2011). *Handbook of semantic Web technologies. Vol. 1, Foundations and technologies.* Berlin: Springer.
45. Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5, 199-220.
46. Slimani, T. (2015). A Study Investigating Typical Concepts and Guidelines for Ontology Building. *Journal of Emerging Trends in Computing and Information Sciences, 5*(12), 886-893.

47. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J. & Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. In *6th International Semantic Web Conference,* (p./pp. 11-15), Busan, Korea.
48. Passant, A. (2010). Measuring Semantic Distance on Linking Data and Using it for Resources Recommendations. *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence, 77,* (p./pp. 1123), Stanford, California, USA
49. Hoser, B., Hotho, A., Jäschke, R., Schmitz, C. & Stumme, G. (2006). Semantic Network Analysis of Ontologies. *Proceedings of the 3rd European Semantic Web Conference, 4011,* (p./pp. 514-529), June, Budva, Montenegro.
50. Finin, T. W., Ding, L., Pan, R., Joshi, A., Kolari, P., Java, A. & Peng, Y. (2005). Swoogle: Searching for Knowledge on the Semantic Web. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management,* (p./pp. 1682-1683), New York, NY, USA.
51. Cheng, G. & Qu, Y. (2008). Term Dependence on the Semantic Web. In *Proceedings of 7th International Semantic Web Conference (ISWC)*, *5318,* (p./pp.665–680), Karlsruhe, Germany.
52. Hausenblas, M., Halb, W., Raimond, Y. & Heath, T. (2008). What is the Size of the Semantic Web? *I-Semantics 2008: International Conference on Semantic Systems, 5318,* (p./pp. 665-680), Graz, Austria.
53. Rodriguez, M. A. (2009). A Graph Analysis of the Linked Data Cloud. *CoRR, abs/0903.0194.*
54. Auer, S., Demter, J., Martin, M. & Lehmann, J. (2012). LODStats - An Extensible Framework for High-Performance Dataset Analytics. In *A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. d'Aquin, A. Nikolov, N. Aussenac-Gilles & N. Hernandez (eds.), EKAW, 7603,* (p./pp. 353-362), Galway City, Ireland.
55. Dividino, R., Scherp, A., Gröner, G. & Gottron, T. (2013). Change-a-LOD: Does the Schema on the Linked Data Cloud Change or Not? In *COLD'13: International Workshop on Consuming Linked Data, 1034*, (p./pp. 87-98), Sydney, Australia.
56. Schmachtenberg, M., Bizer, C. & Paulheim, H. (2014). Adoption of the Linked Data Best Practices in Different Topical Domains. In *P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. A. Knoblock, D. Vrandecic, P. T. Groth, N. F. Noy, K. Janowicz & C. A. Goble (eds.), Semantic Web Conference, (1),* (p./pp. 245-260).
57. Morgan, L.H. (1851). *League of the Hodénosaunee or Iroquois*. Rochester, NY: Sage.
58. Morgan, L. H. (1851). *Ancient society*.
59. Macfarlane, A. (1883). Analysis of Relationships of Consanguinity and Affinity. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 12, 46.
60. Hobson, J. (1884). The Evolution of Modern Capitalism: A Study of Machine Production. *Journal of The Royal Statistical Society*, *69*(4), 780.

61. Euler, L. (1736). *Mechanica, sive, Motus scientia analytice exposita*. Petropoli: Ex typographia Academiæ Scientiarum.
62. Scott, J. (2000). *Social Network Analysis: A Handbook*. Sage Publications.
63. Wasserman, S., Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge University Press.
64. Gansner, E.R. & North, S.C. (2000). An open graph visualization system and its applications to software engineering. *Software: Practice and Experience*, 30(11), 1203-1233.
65. Peixoto, T. P. (2014). The graph-tool python library.
66. Freeman, L. C. (2004). *The Development of Social Network Analysis: A Study in the Sociology of Science*. *Contemporary Sociology: A Journal of Reviews, 37*(3), 221-222.
67. Hagberg, A. A., Schult, D. A. & Swart, P. J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. In *G. Varoquaux, T. Vaught & J. Millman (eds.), Proceedings of the 7th Python in Science Conference*, (p./pp. 11-15), Pasadena, CA.
68. Smith, M., Milic-Frayling, N., Shneiderman, B., Mendes Rodrigues, E., Leskovec, J., & Dunne, C., (2010). NodeXL: a free and open network overview, discovery and exploration add-in for Excel 2007/2010.
69. Batagelj, V. & Mrvar, A. (2003). Pajek - analysis and visualization of large networks. *Graph Drawing Software. M. Juenger and P. Mutze., Graph Drawing. GD 2001. Lecture Notes in Computer Science, (2265),* (pp. 77-103).
70. Bastian, M., Heymann, S. & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. In E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov & B. L. Tseng (eds.), *ICWSM*: The AAAI Press, (8), (p./pp. 361-362)
71. Kelz, A., & Hodic, A. (2014). Facebook and the changing way We speak. In *Proceedings of the European Conference on Social Media: ECSM 2014, Academic Conferences Limited*, (p./pp. 249)
72. Khoshnood, F. (2012). Designing a Recommender System Based on Social Networks and Location Based Services. *International Journal of Managing Information Technology*, *4*(4), 41-47.
73. Srivastav, M.K. & Nath, A. (2015). Mathematical Modelling of Mutual Relationship and Countable Extension of Connected Nodes in Social Networking. In *International Journal International Journal of Advance Research in Computer Science and Management Studies, (3)*5, (p./pp. 1-6)
74. Schaeffer, S. (2007). Graph clustering. *Computer Science Review*, *1*(1), 27-64.
75. Moreno, J. (1932). The Application of the Group Method to the Classification of Prisoners. *Sociometry*, *8*(3/4), 15.

76. Moreno, J. (1934). Sociometric Theory of Leadership and Isolation in Who Shall Survive? *Sociometry*, *13*(4), 382.
77. Moreno, J., & Jennings, H. (1938). Statistics of Social Configurations. *Sociometry*, *1*(3/4), 342.
78. Handlin, O., Warner, W., & Lunt, P. (1942). The Social Life of a Modern Community. *The New England Quarterly*, *15*(3), 554.
79. Mayo, E. (1946). The Social Problems of an Industrial Civilization. *Harvard Law Review*, *59*(5), 830.
80. Lewin, K., & Lippitt, R. (1938). An Experimental Approach to the Study of Autocracy and Democracy: A Preliminary Note. *Sociometry*, *1*(3/4), 292.
81. Tsalatsanis, A., Barnes, L., Hozo, I., Skvoretz, J., & Djulbegovic, B. (2011). A Social Network Analysis of Treatment Discoveries in Cancer. *Plos ONE*, *6*(3), e18060.
82. Novielli, N. & Marczak, S. (2013). Social Network Analysis for Global Software Engineering: Exploring Developer Relationships from a Fine-Grained Perspective. In *IEEE 8th International Conference on Global Software Engineering Workshops (ICGSEW), (*p./pp. 47-48), Bari, Italy.
83. Koochakzadeh, N., Kianmehr, K., Sarraf, A. & Alhajj, R. (2012). Stock Market Investment Advice: A Social Network Approach. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012),* (p./pp. 71-78).
84. Grunspan, D., Wiggins, B., & Goodreau, S. (2014). Understanding Classrooms through Social Network Analysis: A Primer for Social Network Analysis in Education Research. *CBE—Life Sciences Education*, *13*(2), 167-178.
85. Valente, T., Palinkas, L., Czaja, S., Chu, K., & Brown, C. (2015). Social Network Analysis for Program Implementation. *PLOS ONE*, *10*(6), e0131712.
86. Pattison, P. (1993). *Algebraic Model for Social networks.* Cambridge (MA): Cambridge University Press.
87. Kannan, S., Khanna, S. & Roy, S. (2008). STCON in Directed Unique-Path Graphs. In *R. Hariharan, M. Mukund & V. Vinay (eds.), International Conference on the Foundations of Software Technology and Theoretical Computer Science (FSTTCS,* (2), (p./pp. 256-267)
88. Passant, A. (2010). Measuring Semantic Distance on Linking Data and Using it for Resources Recommendations. In *AAAI Spring Symposium: Linked Data Meets Artificial Intelligence*, (77), (p./pp. 123)
89. Feng, X., Chang, L., Lin, X., Qin, L. & Zhang, W. (2016). Computing Connected Components with linear communication cost in pregel-like systems. In IEEE 32nd International Conference on Data Engineering (ICDE)*,* (p./pp. 85-96)
90. Tarjan, R. (1972). Depth-First Search and Linear Graph Algorithms. *SIAM Journal on Computing*, *1*(2), 146-160.

91. Nuutila, E. & Soisalon-Soininen, E. (1994). On Finding the Strongly Connected Components in a Directed Graph. *Journal of Information. Process Letter.* (49), 9-14.
92. Tauro, S. L., Palmer, C. R., Siganos, G. & Faloutsos, M. (2001). A simple conceptual model for the Internet topology. In *Global Telecommunications Conference, GLOBECOM,* (p./pp. 1667-1671), San Antonio, Texas, USA.
93. Opsahl, T., Agneessens, F. & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks*. (32), 245-251.
94. Freeman, L. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, *40*(1), 35.
95. Ehrig, M. & Euzenat, J. (2005). Relaxed Precision and Recall for Ontology Matching. In *B. Ashpole, M. Ehrig, J. Euzenat & H. Stuckenschmidt (eds.), Proceedings of the K-CAP 2005 Workshop on Integrating Ontologies*, (156), (p./pp. 25–32), Banff, Canada.
96. Euzenat, J., & Shvaiko, P. (2013). *Ontology Matching*. Springer, Berlin, Heidelberg.
97. Giunchiglia, F. & Shvaiko, P. (2003). Semantic matching. *The Knowledge Engineering Review*, *18*(3), 265-280.
98. Tönnies, F. (1925). Gemeinschaft und Gesellschaft. (Theorem der Kultur-Philosophie.). *Kant-Studien*, *30*(1-2).
99. Faria, D., Pesquita, C., Santos, E., Palmonari, M., Cruz, I. F., & Couto, F. M. (2013). The AgreementMakerLight ontology matching system. In *R. Meersman, H. Panetto, T. S. Dillon, J. Eder, Z. Bellahsene, N. Ritter, P. D. Leenheer & D. Dou (eds.), Confederated International Conferences on On the Move to Meaningful Internet Systems,* (8185), (p./pp. 527-541), Graz, Austria.
100. Jiménez-Ruiz, E. & Grau, B. C. (2011). LogMap: Logic-Based and Scalable Ontology Matching. In *L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. F. Noy & E. Blomqvist (eds.), International Semantic Web Conference, (7031),* (p./pp. 273-288), Bonn, Germany.
101. Djeddi, W., & Khadir, M. (2013). Ontology alignment using artificial neural network for large-scale ontologies. *International Journal of Metadata, Semantics and Ontologies*, *8*(1), 75.
102. Khiat, A. (2016). CroLOM: cross-lingual ontology matching system results for OAEI 2016. In *P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, M. Cheatham, O. Hassanzadeh & R. Ichise (eds.), Proceedings of the 11th International Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016),* (7031), (p./pp. 146-152) Kobe, Japan.
103. Rybinski, M., del Mar Roldán García, M., García-Nieto, J. & Montes, J. F. A. (2016). DisMatch results for OAEI 2016. In *P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, M. Cheatham, O. Hassanzadeh & R. Ichise (eds.), Proceedings of the 11th International Workshop on Ontology*

Matching co-located with the 15th International Semantic Web Conference (ISWC 2016), (7031), (p./pp. 161-165), Kobe, Japan.

104. Rodríguez-García, M., Gkoutos, G., Schofield, P., & Hoehndorf, R. (2016). Integrating phenotype ontologies with PhenomeNET. *Journal of Biomedical Semantics*, *8*(1).

105. Megdiche, I., Teste, O. & dos Santos, C. T. (2016). An Extensible Linear Approach for Holistic Ontology Matching. In P. T. Groth, E. Simperl, A. J. G. Gray, M. Sabou, M. Krötzsch, F. Lécué, F. Flöck & Y. Gil (eds.), *The Semantic Web – ISWC 2016. Lecture Notes in Computer Science, (9981),* (p./pp. 393-410)

106. Zhao, M. & Zhang, S. (2016). FCA-Map results for OAEI 2016. In *P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, M. Cheatham, O. Hassanzadeh & R. Ichise (eds.), Proceedings of the 11th International Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016),* (7031), (p./pp. 172-177), Kobe, Japan.

107. da Silva, J., Baião, F. A. & Revoredo, K. (2016). ALIN results for OAEI 2016. In *P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, M. Cheatham, O. Hassanzadeh & R. Ichise (eds.), Proceedings of the 11th International Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016),* (7031), (p./pp. 130-137), Kobe, Japan.

108. Thanh, T. D. (2011). *Process-oriented Semantic Web Search.* Unpublished doctoral dissertation, Karlsruher Institut für Technologie (KIT), Fakultät für Wirtschaftswissenschaften (Fak. f. Wirtschaftswiss.) Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB)

109. van Rijsbergen, C. J. (1979). *Information retrieval*. London: Butterworths.

110. Butt, A. S., Haller, A., and Xie, L. (2015). A taxonomy of semantic web data retrieval techniques. In *Proceedings of the 8th International Conference on Knowledge Capture,* (9), (p./pp. 1-9).

111. de Maele, F. V., Spyns, P. & Meersman, R. (2008). An Ontology-Based Crawler for the Semantic Web. In *R. Meersman, Z. Tari & P. Herrero (eds.), On the Move to Meaningful Internet Systems: OTM 2008 Workshops,* (5333), (p./pp. 1056-1065)

112. Isele, R., Harth, A., Umbrich, J. & Bizer, C. (2010). LDspider: An open-source crawling framework for the Web of Linked Data. In *A. Polleres & H. Chen (eds.), poster at the International Semantic Web Conference (ISWC2010), (658),* (p./pp. 29-32), Shanghai, China.

113. Haarslev, V. & Moller, R. (2003). Racer: An OWL reasoning agent for the semantic web. In *Proceedings of the International Workshop on Applications, Products and Services of Web-Based Support Systems, in Conjunction with 2003 IEEE/WIC International Conference on Web Intelligence,* (p./pp. 27–32), Sanibel Island, Florida, USA.

114. Glimm, B., Horrocks, I., Motik, B., Stoilos, G., & Wang, Z. (2014). HermiT: An OWL 2 Reasoner. *Journal of Automated Reasoning*, *53*(3), 45-269.

115. Cantador, M. F. I. & Castells, P. (2007). Improving Ontology Recommendation and Reuse in WebCORE by Collaborative Assessments. In *Workshop on Social and Collaborative Construction of Structured Knowledge (CKC 2007) at WWW 2007*, (273)*,* (p./pp. 145–148), Banff, Canada.

116. Tran, P. N. & Nguyen, D. T. (2016). A Linked Data Driven Semantic Model for Interpreting English Queries in Question Answering System. In *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*, (21), (p./pp. 1-5), Danang, Vietnam.

117. Lux, M., zu Eissen, S. M. & Granitzer, M. (2006). Graph Retrieval with the Suffix Tree Model. In *Proceedings of the ECAI'06 3rd International Workshop on Text-based Information Retrieval (TIR-06)*, (p./pp. 30-34), Trento, Italy.

118. He, H., Wang, H., Yang, J. & Yu, P. S. (2007). BLINKS: ranked keyword searches on graphs. *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data,* (p./pp. 305--316), Beijing, China.

119. Yuan, D., & Mitra, P. (2012). Lindex: a lattice-based index for graph databases. *The VLDB Journal*, *22*(2), 229-252.

120. Noy, N. F., Shah, N. H., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Montegut, M. J., Rubin, D. L., Youn, C. & Musen, M. A. (2008). BioPortal: A Web Repository for Biomedical Ontologies and Data Resources. In *Proceedings of ISWC 2008*. *International Semantic Web Conference (Posters & Demos).*

121. Vesse, R., Hall, W. & Carr, L. (2010). Preserving Linked Data on the Semantic Web by the application of Link Integrity techniques from Hypermedia. In C. Bizer, T. Heath, T. Berners-Lee & M. Hausenblas (eds.), *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010,* Raleigh, USA.

122. Guo, Y. & Heflin, J. (2007). Document-Centric Query Answering for the Semantic Web. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, (p./pp. 409-415)

123. Dietze, S. (2016). Retrieval, Crawling and Fusion of Entity-centric Data on the Web. In A. Calì, D. Gorgan & M. Ugarte (eds.), *International KEYSTONE Conference,* (p./pp. 3-16), Fremont, CA, USA.

124. Cheng, G., Zhang, Y. & Qu, Y. (2014). Explass: Exploring Associations between Entities via Top-K Ontological Patterns and Facets. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. A. Knoblock, D. Vrandecic, P. T. Groth, N. F. Noy, K. Janowicz & C. A. Goble (eds.), *International Semantic Web Conference, (2),* (p./pp. 422-437)

125. Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., & Sahin, S. et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biology*, *16*(1), 22.

126. Hogan, A., Harth, A., Umbrich, J., Kinsella, S., Polleres, A., & Decker, S. (2011). Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine. *Web Semantics: Science, Services and Agents on The World Wide Web*, *9*(4), 365-401.

127. Emamdadi, R., Kahani, M., & Zarrinkalam, F. (2014). A focused linked data crawler based on HTML link analysis. In *4th International Conference on Computer and Knowledge Engineering (ICCKE)*, (p./pp. 74-79), Mashhad, Iran.

128. Jain, N. & Rawat, P. (2013). A Study of Focused Web Crawlers for Semantic Web, *International Journal of Computer Science and Information Technologies, 4*(3), 398-402.

129. Furche, T., Olteanu, D., & Steinhaus, R. (2010). *G-Store: A Storage Manager for Graph Data.* Ph.D. dissertation, Citeseer.

130. Cudré-Mauroux, P., Enchev, I., Fundatureanu, S., Groth, P. T., Haque, A., Harth, A., Keppmann, F. L., Miranker, D. P., Sequeda, J. & Wylot, M. (2013). NoSQL Databases for RDF: An Empirical Evaluation. In *H. Alani, L. Kagal, A. Fokoue, P. T. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. F. Noy, C. Welty & K. Janowicz (eds.), International Semantic Web Conference, (2),* (p./pp. 310-325), Sydney, NSW, Australia.

131. Minack, E., Sauermann, L., Grimnes, G., Fluit, C. & Broekstra, J. (2008). *The Sesame Lucene Sail: RDF Queries with Full-text Search.* (Technical Report2008-1). NEPOMUK Consortium.

132. Fox, G.C., Lee, M.S., Nguyen, M.D., & Oh, S. (2014). SPARQL Query Optimization for Structural Indexed RDF Data.

133. Udrea, O., Pugliese, A. & Subrahmanian, V. S. (2007). GRIN: A Graph Based RDF Index. In *Proceedings of the 21nd AAAI Conference on Artificial Intelligence (AAAI),* (2), (p./pp. 1465-1470), Vancouver, British Columbia, Canada.

134. Sankar, S., Singh, M., Sayed, A., & Alkhalaf Bani-Younis, J. (2014). An Efficient and Scalable RDF Indexing Strategy based on B-Hashed-Bitmap Algorithm using CUDA. *International Journal of Computer Applications*, *104*(7), 31-38.

135. Klusch, M., Fries, B. & Sycara, K. (2006). Automated semantic web service discovery with OWLS-MX. In *H. Nakashima, M. P. Wellman, G. Weiss & P. Stone (eds.), Conference: 5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006),* (p./pp. 915-922), Hakodate, Japan.

136. Alani, H., Brewster, C. & Shadbolt, N. (2006). Ranking Ontologies with AKTiveRank. In *I. F. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold & L. Aroyo (eds.), International Semantic Web Conference,* (4273), (p./pp. 1-15), Athens, GA, USA.

137. Ning, X., Jin, H., & Wu, H. (2008). RSS: A framework enabling ranked search on the semantic web. *Information Processing & Management*, *44*(2), 893-909.

138. Lamberti, F., Sanna, A., & Demartini, C. (2009). A Relation-Based Page Rank Algorithm for Semantic Web Search Engines. *IEEE Transactions on Knowledge and Data Engineering*, *21*(1), 123-136.

139. Franz, T., Schultz, A., Sizov, S. & Staab, S. (2009). TripleRank: Ranking Semantic Web Data by Tensor Decomposition. In *Bernstein A. et al. (eds) The Semantic Web - ISWC 2009. ISWC 2009. Lecture Notes in Computer Science,* (5823), (p./pp. 213--228), Chantilly, VA, USA.

140. Meymandpour, R. & Davis, J. G. (2013). Linked Data Informativeness. In *Y. Ishikawa, J. Li, W. Wang, R. Zhang & W. Zhang (eds.), Web Technologies and Applications. APWeb 2013. Lecture Notes in Computer Science,* (7808), (p./pp. 629-637), Sydney, Australia.

141. Fellbaum, C. (1998). *WordNet*. Cambridge, Mass.: Massachusetts Institute of Technology.

142. Turney, P. D. & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37, 141-188.

143. Butt, A. S., Haller, A. & Xie, L. (2014). Relationship-Based Top-K Concept Retrieval for Ontology Search. In *Janowicz K., Schlobach S., Lambrix P., Hyvönen E. (eds) Knowledge Engineering and Knowledge Management. EKAW 2014. Lecture Notes in Computer Science,* (8876), (p./pp. 485-502).

144. Zhang, X., Cheng, G. & Qu, Y. (2007). Ontology summarization based on rdf sentence graph. In *Proceedings of the 16th international conference on World Wide Web*, (p./pp. 707--716), Banff, Alberta, Canada

145. Alahmari, F., Magee, L., & Thom, J.A. (2014). A model for ranking entity attributes using DBpedia. *Aslib Journal of Information Management, 66*(5), 473-493.

146. Pérez-Agüera, J. R., Arroyo, J., Greenberg, J., Perez-Iglesias, J. & Fresno, V. (2010). Using BM25F for Semantic Search. In *Proceedings of the 3rd International Semantic Search Workshop, SEMSEARCH '10,* (2)*,* (p./pp. 1–8), North Carolina, USA.

147. Arora, P., & Vikas, O. (2012). Semantic Searching and Ranking of Documents using Hybrid Learning System and WordNet. *International Journal of Advanced Computer Science and Applications*, *3*(6)

148. Roatis, A. (2014). *Efficient Querying and Analytics of Semantic Web Data. (Interrogation et Analyse Efficiente des Données du Web Sémantique).* Unpublished doctoral dissertation, University of Paris-Sud, Orsay, France.

149. Li, Y., Qasem, A. & Heflin, J. (2010). A Scalable Indexing Mechanism for Ontology-Based Information Integration. In *J. X. Huang, I. King, V. V. Raghavan & S. Rueger (eds.), Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International,* (01), (p./pp. 328-331), Toronto, ON, Canada.

150. Liang, Y., Wang, H., Liu, Q., Tran, T., Penin, T. & Yu, Y. (2008). Efficient Index Maintenance for Frequently Updated Semantic Data. In *J. Domingue & C. Anutariya (eds.), The Semantic Web. ASWC 2008. Lecture Notes in Computer Science, (5367),* (p./pp. 182-196), Bangkok, Thailand.

151. Sappagh, S., & Elmogy, M. (2016). A Decision Support System for Diabetes Mellitus Management. *Diabetes Case Reports*, *01*(01).

152. Schiessl, M. & Bräscher, M. (2017). Ontology lexicalization: Relationship between content and meaning in the context of Information Retrieval. *Transinformação*, *29*(1), 57-72.

153. Yazhmozhi, V. M., Ramya Soundarya Lakshmi, K. & Sreessruthi, S. (2013)**.** Semantically Classified Web Portal for Engineers. *International Journal of Engineering Research & Technology, 2*(3).

154. Sun H., Weng J., Yu G. & Massawe R. H. (2013). *A DNA-Based Semantic Fusion Model for Remote Sensing Data. PLoS ONE 8*(10): e77090.

155. Schmidt, M., Görlitz, O., Haase, P., Ladwig, G., Schwarte, A. & Tran, T. (2011). FedBench: A Benchmark Suite for Federated Semantic Data Query Processing. In *L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. F. Noy & E. Blomqvist (eds.), International Semantic Web Conference, (7031),* (p./pp. 585-600)

156. Fafalios, P., Yannakis, T. & Tzitzikas, Y. (2016). Querying the Web of Data with SPARQL-LD. In *Fuhr N., Kovács L., Risse T., Nejdl W. (eds) Research and Advanced Technology for Digital Libraries, TPDL 2016, Lecture Notes in Computer Science,* (9819), (p./pp. 175-187)

157. Zhang, L., Zhu, M., & Huang, W. (2009). A Framework for an Ontology-based E-commerce Product Information Retrieval System. *Journal of Computers*, *4*(6).

158. Chen, L., Mulvenna, M., & Nugent, C. (2008). *Semantic Smart Homes: Towards Knowledge Rich Assisted Living EnvironmentsIntelligent Patient Management*. Studies in Computational Intelligence, (189), Springer, Berlin, Heidelberg.

159. Mika, P. & Potter, T. (2012). Metadata Statistics for a Large Web Corpus. In C. Bizer, T. Heath, T. Berners-Lee & M. Hausenblas (eds.), WWW2012 Workshop on Linked Data on the Web LDOW, (937), Lyon, France.

160. David, Jé., Euzenat, Jé., Scharffe, F. & Trojahn dos Santos, Cá. (2011). The Alignment API 4.0. *Semantic Web Journal, 2*(1), 3-10.

161. Nogales, A., Sicilia, M.-Á., Alonso, S. S. & Barriocanal, E. G. (2016). Linking from Schema.org microdata to the Web of Linked Data: An empirical assessment. *Computer Standards & Interfaces*, *45*, 90-99.

162. Nogales, A., Sicilia, M.-Á., Barriocanal, E. G. & Alonso, S. S. (2013). Exploring the Potential for Mapping Schema.org Microdata and the Web of Linked Data. In *E. Garoufallou & J. Greenberg (eds.), Metadata and Semantics Research. MTSR 2013. Communications in Computer and Information Science,* (390), (p./pp. 266-276)

163. Mühleisen, H. & Bizer, C. (2012). Web Data Commons - Extracting Structured Data from Two Large Web Corpora. In *C. Bizer, T. Heath, T. Berners-Lee & M. Hausenblas (eds.), Proceedings of the WWW2012, Workshop on Linked Data on the Web (LDOW2012),* (937), (p./pp. 133-145), Lyon, France.

164. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, Sö. & Bizer, C. (2015). DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 6, 167-195.

165. Krafft, D. B., Cappadona, N. A., Caruso, B., Corson-Rikert, J., Devare, M., Lowe, B. J. & Collaboration, V. (2010). VIVO: Enabling National Networking of Scientists. In *Proceedings of the Web Science Conference*.

166. Lezcano, L., Jörg, B. & Sicilia, M.A. (2012) Modeling the Context of Scientific Information: Mapping VIVO and CERIF**.** In *Bajec M., Eder J. (eds) Advanced Information Systems Engineering Workshops. CAiSE 2012. Lecture Notes in Business Information Processing,* 112, (p./pp. 123-129), Gdańsk, Poland.

167. Lezcano, L., Jörg, B., Lowe, B. & Corson-Rikert, J. (2013). Promoting International Interoperability of Research Information Systems: VIVO and CERIF. *Journal of Universal Computer Science, 19*, 1854-1867.

168. Nogales, A., Sicilia, M.-Á. & Jörg, B. (2014). Combining VIVO and Google Scholar Data as Sources for CERIF Linked Data: A Case in the Agricultural Domain. In *K. G. Jeffery, A. Clements, P. de Castro & D. Luzi (eds.), CRIS,* (p./pp. 266-271)

169. Nogales, A., Sicilia-Urban, M.A. & García-Barriocanal, E. (2017). Measuring vocabulary use in the Linked Data Cloud. *Online Information Review, 41*(2), 252-271.